Data Science Research Project A in the School of Mathematical Sciences

Ruyu Liu
A1882974

November 22, 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**

Abstract

Hybrid communication systems have become increasingly popular and widely applied across various domains. This study focuses on developing predictive models for signal attenuation in Free Space Optical (FSO) and Radio Frequency (RF) channels using synthetic empirical data from hybrid communication systems. Two machine learning approaches, Random Forest and Neural Networks, were employed to construct and evaluate models for each channel. The Random Forest models included both generic (GRF) and specific (SRF) models to account for varying weather conditions, while Neural Networks explored Multi-Layer Perceptrons (MLPs) and Convolutional Neural Networks (CNNs).

During the modelling process, data preprocessing techniques were applied, including feature selection for Random Forest models and One-Hot Encoding with Standardised Scaling for Neural Network models. Hyper-parameter tuning was performed to optimise all models.

Comparative evaluations on the validation set identified the GRF model as the best-performing model for both FSO and RF channels. Final evaluations on the test set confirmed its strong performance, achieving an RMSE of 0.82 dB and an R^2 value of 0.96 for the FSO channel and an RMSE of 0.49 dB and an R^2 value of 0.98 for the RF channel.

Future research will investigate the correlation between FSO and RF channels, providing deeper insights into their responses to similar environmental conditions and supporting the development of robust hybrid communication systems.

1 Introduction

In the field of wireless communication, Radio Frequency (RF) has been the most commonly used method for data transmission in its early stages. However, RF channels struggle to meet the increasing demand for higher data rates. To address this limitation, Free Space Optical (FSO) communication, which uses the power of optical signals for data transmission, has emerged as a promising alternative. Despite its advantages, FSO communication is highly sensitive to environmental factors, particularly weather conditions[18]. To overcome these limitations, a hybrid Radio Frequency/Free Space Optical (RF/FSO) system has been developed, combining both communication channels to enable stable data transmission under diverse and challenging outdoor environments[3]. This hybrid system has found applications across various fields, including Smart Cities, Healthcare, and Deep Space Communication[14, 1].

To assess the performance of the hybrid RF/FSO system and understand the impact of different weather conditions, it is essential to develop predictive models for signal attenuation. These models predict the degradation in signal strength and quality under different environmental conditions during data transmission. While traditional models, such as those recommended by ITU-P and its extensions, are widely used for specific weather conditions and applications[16], their reliance on extensive weather parameter measurements limits their applicability. In contrast, more general machine learning approaches have emerged as a powerful alternative.

Among supervised machine learning methods, each algorithm presents distinct advantages and limitations. Linear regression is effective for modelling linear relationships but struggles with non-linear data. Random Forest (RF) models are interpretable, computationally efficient, and require minimal hyper-parameter tuning, yet they are prone to over- fitting and may not effectively capture inter-feature correlations. Neural Networks (NN), excel in capturing complex patterns, though they are computationally intensive, require significant training time, and can be challenging to interpret.

This study explores two machine learning approaches—Random Forest and Neural Networks —to develop predictive models for attenuation in FSO and RF channels. Both approaches follow the same methodology for model development, and their performance is evaluated using metrics such as Root Mean Squared Error (RMSE) and R-squared value (R^2).

The remainder of this paper is structured as follows:

- **Background:** An overview of the impact of weather conditions on RF/FSO communication and a description of the dataset and its properties.
- **Methods:** A detailed explanation of the methodologies employed, including Random Forest and Neural Network models, evaluation metrics, and the overall workflow.
- **Results:** Key findings, including model performance comparisons and the final selected models for each channel.
- **Conclusion:** A summary of the study's contributions, along with directions for future research.

2 Background

2.1 Properties of RF and FSO systems

The origins of Radio Frequency (RF) communication can be traced back to Maxwell's discovery of electromagnetic waves, which Heinrich Hertz later confirmed experimentally [4]. RF communication relies on the propagation of electromagnetic waves to transmit signals. However, signal attenuation occurs during transmission due to factors such as distance and frequency [9]. Additionally, varying weather conditions significantly impact RF signal attenuation. For instance, in rainy conditions, the scattering and absorption of electromagnetic waves by water droplets result in rain attenuation [22]. Similarly, in sandstorms, the presence of sand and dust particles in the air leads to attenuation and depolarisation of electromagnetic waves upon interaction with these particles [21].

In contrast, Free Space Optical (FSO) systems are even more sensitive to weather conditions. Rain and fog significantly degrade the performance of FSO systems. Suspended water and dust particles cause attenuation and scattering of the light beam, leading to signal degradation, reduced link range, and increased error rates [11]. Furthermore, wind and temperature fluctuations also adversely affect optical signal strength. High temperatures and wind speeds induce turbulence, resulting in scintillation, which causes fluctuations in signal intensity [2].

2.2 Properties of dataset

The dataset used in this project is real synthetic empirical data derived from a hybrid communication system operating in six cities worldwide. It comprises a total of 91,379 measurements and includes two output variables: Radio Frequency (RF) channel attenuation and Free Space Optical (FSO) channel attenuation. Additionally, it includes 25 weather parameters as input features, such as humidity, visibility, and wind speed.

For convenience and clarity of analysis, each feature is assigned a specific abbreviation. The abbreviations consist of a prefix, representing the feature type, and a suffix, indicating the statistical value. Detailed information about each feature type, statistical value, and their corresponding abbreviations is provided in Tab. 1. Features without a suffix represent the median value within the last minute. For example, RI indicates the median rain intensity within the last minute, while RI_Max represents the maximum rain intensity within the last minute.

Table 1: Feature Abbreviations and Their Meanings

(a) Prefix: Feature Type

Abbrv.	**Meaning**
AH	Absolute humidity (g/m^3)
Dist	The link distance (m) between the transmitter and receiver.
Freq	Carrier frequency in the RF system (Hz).
Part	Particle count
RI	Rain intensity (mm/hr)
RH	Relative humidity (%)
SC	Adapted from the surface synoptic observations code.
Temp	Temperature (K)
Hour	The hour in which the measurement was recorded.
Vis	Visibility (m)
WD	The angle (degrees) between the wind direction and the signal propagation path.
WS	Wind speed (m/s)

(b) Suffix: Statistical Value

Abbrv.	**Meaning**
Min	Minimum within the last minute
Max	Maximum within the last minute
Diff	Change within the last minute

The dataset contains no missing or erroneous values. The distribution of the output variables is shown in Fig. 1. FSO attenuation exhibits a right-skewed distribution, with generally low values and occasional extremes, reflecting its sensitivity to environmental conditions. In contrast, RF attenuation follows a stable, normalised distribution, indicating greater robustness under various weather conditions.

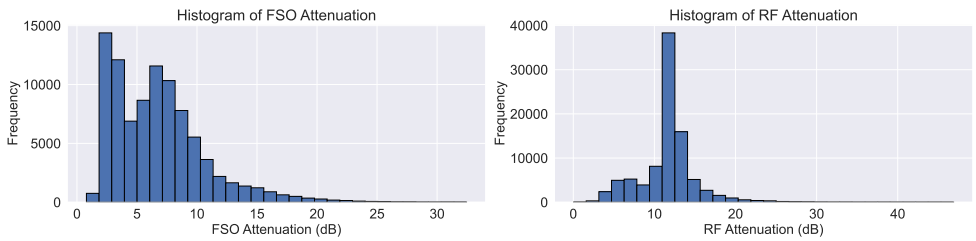


Figure 1: Histograms for Output Variables.

Among the 25 features, three are categorical: Hour, Frequency, and

SC (weather conditions). Bar chart visualisations in Fig. 2 show that the data for Hour and Frequency are relatively balanced. However, SC is significantly imbalanced, with Clear Weather and Rain dominating the dataset, while conditions such as Fog, Snow, and Dust Storms are underrepresented.

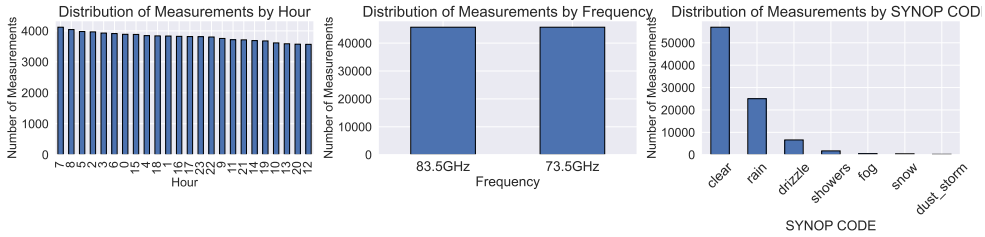
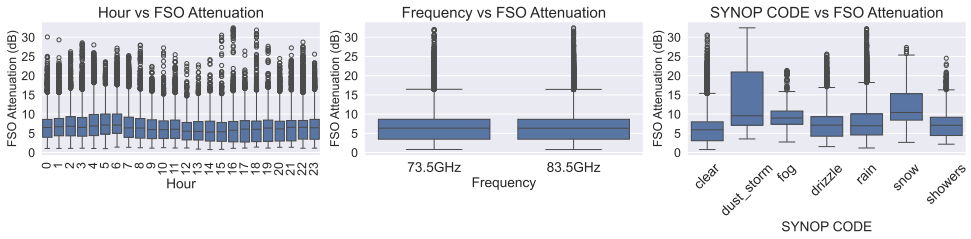


Figure 2: Distributions of Categorical Variables.

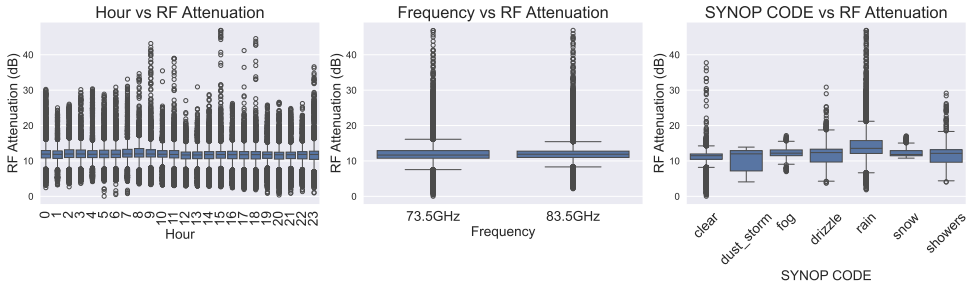
The 22 continuous variables exhibit diverse distributions, as detailed in the Appendices. Absolute Humidity and Temperature display bimodal trends, likely indicating seasonal variations. Particle Count, Rain Intensity, and Wind Speed are right-skewed, with rare extreme values. The Link Distance variable is multi-modal, aligning with standard infrastructure setups.

As shown in Fig. 3, both FSO and RF attenuation demonstrate consistent median values across different hours, indicating less impact from the time of day. Similarly, attenuation is unaffected by frequency, as the distributions for 73.5 GHz and 83.5 GHz are nearly identical. However, weather conditions have a significant influence on both systems. Clear weather results in the lowest and most stable attenuation, while adverse conditions, such as dust storms, rain, and snow, lead to higher attenuation with greater variability. FSO attenuation is particularly sensitive to adverse weather, especially dust storms and rain, whereas RF attenuation remains comparatively stable under similar conditions.

The relationships between continuous variables and FSO/RF attenuation, illustrated through scatter plots in the Appendices, reveal distinct trends. Both FSO and RF attenuation exhibit strong positive correlations with Particle Count and Rain Intensity, as higher values significantly increase attenuation due to scattering and absorption effects. In contrast, variables related to Humidity, Visibility, Temperature, and Wind show weak correlations with both FSO and RF attenuation, indicating limited influence on signal degradation under these conditions.



(a) FSO attenuation versus Hour, Frequency, and SYNOP Code.



(b) RF attenuation versus Hour, Frequency, and SYNOP Code.

Figure 3: Box-plots showing the relationship between categorical variables (Hour, Frequency, SYNOP Code) and attenuation. (a) FSO attenuation. (b) RF attenuation.

Correlations between continuous features, visualised in the Heat map (Fig. 4), reveal strong interdependence within certain feature groups. Absolute Humidity (AH) and its derivatives (AH_Max, AH_Min) demonstrate perfect correlations (1.0), as do Temperature (Temp) and its maximum and minimum values. Additionally, Rain Intensity (RI) and Particle Count (Part) exhibit moderate positive correlations (0.65), suggesting that higher particle counts are associated with increased rain intensity. These findings indicate potential redundancy in some features, and the need for dimensionality reduction before model development. This is particularly important for Random Forest models, which are highly influenced by strongly correlated features.

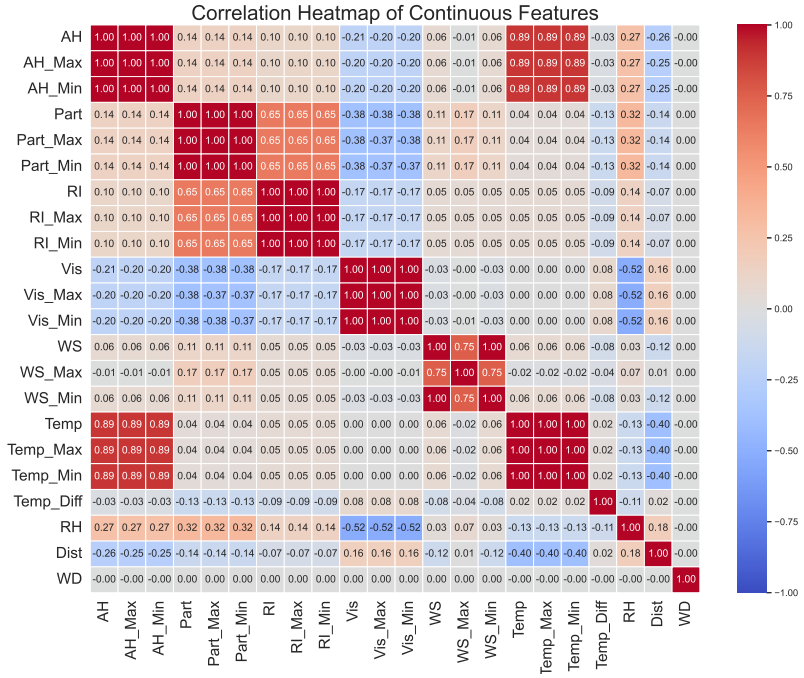


Figure 4: Correlation Heat map of Continuous Features

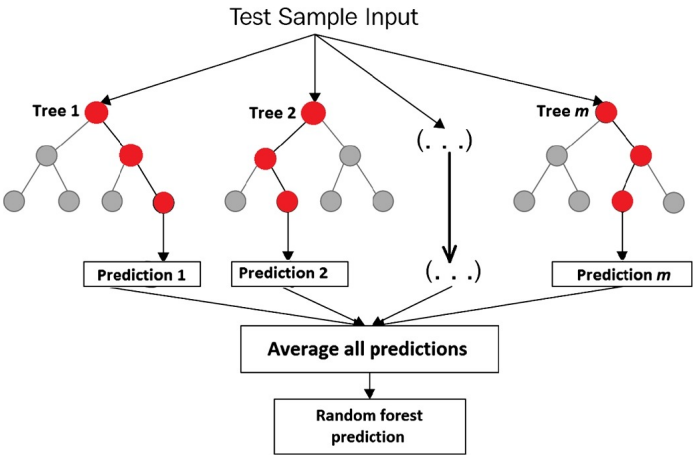
While some extreme values exist in the dataset, these are meteorologically reasonable and reflect real-world conditions (e.g., Dust Storms and Showers). As such, these values are retained rather than treated as errors. To address skewed distributions, preprocessing steps like scaling and feature transformation will be applied in neural network modelling.

3 Methods

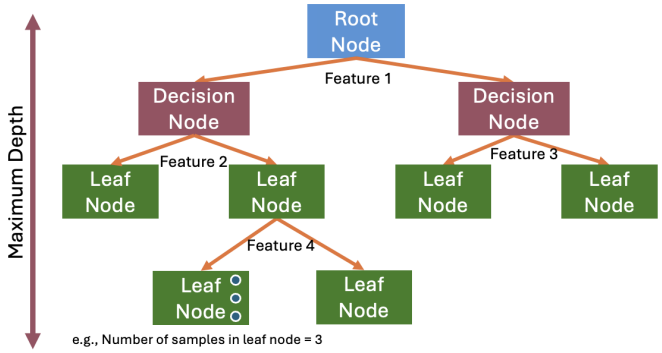
This section describes the main concepts and methodologies employed in developing the Random Forest and Neural Network models. It also outlines the evaluation metrics used to assess each model's performance. Lastly, the overall experimental workflow is presented to provide a comprehensive understanding of the approach.

3.1 Random Forest

Random Forest is an ensemble learning algorithm that enhances prediction accuracy by combining multiple decision trees [20, 15]. In regression tasks, it aggregates predictions from all trees by averaging their outputs, resulting in a robust and reliable model, as illustrated in Fig. 5a [6, 7].



(a) Random Forest. Adapted from [15].



(b) Decision Tree.

Figure 5: The structure of (a) Random Forest and (b) Decision Tree.

3.1.1 Generic and Specific model

Both FSO and RF systems are significantly influenced by weather conditions during signal transmission. To account for this, two approaches were explored when constructing the Random Forest model: a generic model, where the weather condition feature (SYNOP code) is treated as a categorical variable, and specific models, where separate Random Forest models are developed for each of the seven weather conditions represented by the SYNOP code.

Table 2: Random Forest Models and Weather Conditions

Model	Weather Condition	# of Features
Generic Model	All Weather Conditions	FSO: 24, RF: 25
Specific Models	Clear (SC=0)	FSO: 24, RF: 25
	Dust Storm (SC=3)	FSO: 24, RF: 25
	Fog (SC=4)	FSO: 24, RF: 25
	Drizzle (SC=5)	FSO: 24, RF: 25
	Rain (SC=6)	FSO: 24, RF: 25
	Snow (SC=7)	FSO: 24, RF: 25
	Showers (SC=8)	FSO: 24, RF: 25

A total of 14 models were initialised: one generic model and seven specific models for the FSO system, and one generic model and seven specific models for the RF system. The initial number of features(# of Features) for each model, along with the corresponding weather conditions for the specific models, is summarised in Tab. 2. The performance of the generic model and the specific models was compared under different weather conditions to determine whether the more complex specific models provide better predictive accuracy than the generic model.

3.1.2 Feature Selection

Given the high correlation among features and the susceptibility of the Random Forest model to over-fitting, feature selection was implemented for both generic and specific models prior to training. The feature selection process employed a Wrapper Method with Backward Elimination [13, 10]. This approach begins with all available features and iteratively refines the feature set to identify the most relevant variables.

The specific steps of **Backward Elimination** are as follows:

1. **Starting with All Features:** The process starts by including all available features in the model.
2. **Iterative Removal:** In each iteration:
   - A Random Forest model is trained on the current feature set.
   - Model performance is evaluated using OOB (Out-Of-Bag) information, such as RMSE and R^2, and recorded [19].
   - Feature importance is ranked based on the OOB information, and the least important feature is removed.
   - Note: After each feature removal, the Random Forest model is retrained, as the ranking of the remaining features may change with the updated feature set.
3. **Stopping Criterion:** The process continues until no features remain in the model.

This method allows for monitoring the impact of feature removal on model performance. By examining the sequence of feature eliminations, the turning point in performance change is identified. Features removed before the turning point are classified as unimportant, while those retained at the turning point are deemed important features.

Based on this process, the generic and specific datasets are refined to include only the selected important features. These refined datasets are then used to train the generic and specific Random Forest models, ensuring improved performance and reduced risk of over-fitting.

3.1.3 Hyper-parameters Tuning

Hyper-parameters are pre-defined parameters set before training that directly influence a model's performance [5]. To enhance the performance of the Random Forest models, hyper-parameter optimisation was carried out. Four key hyper-parameters were selected for tuning: number of trees in the forest, maximum depth of each tree, minimum samples required to split a node, and minimum samples required at a leaf node. The optimisation process was conducted using GridSearchCV from the Scikit-learn library, to find the best combination of hyper-parameter values for optimal model performance.

The influence of each hyper-parameter on model performance is summarised as follows:

1. **Number of Trees:** Increasing the number of trees generally improves prediction accuracy by reducing variance. However, this also increases computational cost, and after a certain number of trees, the performance improvement becomes negligible.

2. **Maximum Depth:** This parameter defines the longest path between the root node and a leaf node, as illustrated in Fig. 5b. It controls the growth of each decision tree within the Random Forest, helping to prevent over-fitting by limiting tree complexity.

3. **Minimum Samples Required to Split a Node:** This parameter specifies the minimum number of observations required in a node to allow a split. By increasing this value, the number of splits is reduced, leading to simpler trees and mitigating over-fitting. For example, if the value is 2, meaning any node with more than two observations can be split further.

4. **Minimum Samples Required at a Leaf Node:** This hyperparameter determines the minimum number of samples that must remain in a leaf node after a split. By increasing this value, the growth of the tree is constrained, further reducing over-fitting and ensuring simpler terminal nodes.

3.2 Neural Network

3.2.1 One-Hot Encoding and Standardised Scaling

Neural networks require numerical inputs and are highly sensitive to the scale and distribution of input features. Therefore, data preprocessing is essential before training the neural network models.

Categorical variables were converted into numerical representations using One-Hot Encoding [17]. For example, as shown in Fig. 6, the frequency feature with two categories, 73.5 GHz and 83.5 GHz, was transformed into binary numerical values.

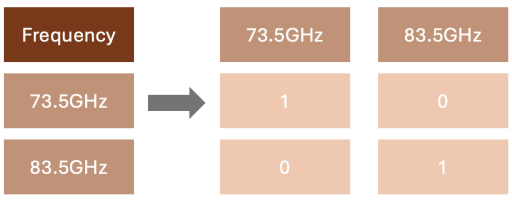


Figure 6: One-Hot Encoding of Frequency Categories.

Some features in the dataset exhibited right-skewed distributions. To address this, standardised scaling was applied prior to building the Neural Network (NN) models, to ensure that each feature had a mean of 0 and a variance of 1, as demonstrated in Fig. 7.

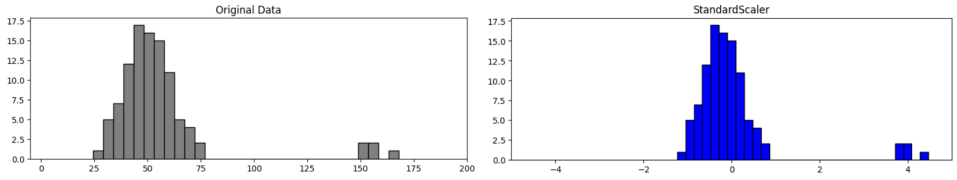
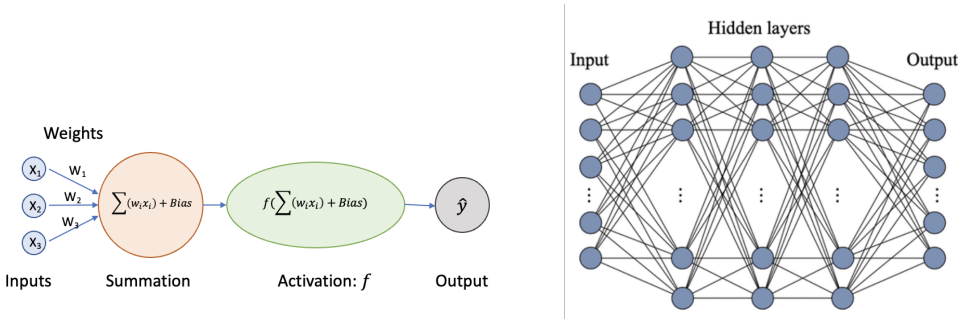


Figure 7: StandardScaler Transformation

3.2.2 Multi-Layer Perceptron

The Perceptron, developed by Frank Rosenblatt in 1958, is a single-layer neural network, as illustrated in Fig. 8a [19]. The Multi-Layer Perceptron (MLP), an extension of the Perceptron, is a fully connected feed-forward neural network consisting of multiple layers of perceptron, each with its own set of weights and activation functions(Fig. 8b).



(a) Single-Layer Perceptron (b) Multi-Layer Perceptron

Figure 8: Single-Layer Perceptron vs Multi-Layer Perceptron

For this regression task, the MLP model employs Mean Squared Error (MSE) as the loss function to measure the error between predicted and actual attenuation values.

$$MSE = \frac{1}{N}\sum_{n=1}^{N}(\hat{y_n} - y_n) \tag{1}$$

where, N represents the number of measurements. $\hat{y_n}$ is the predicted attenuation value and y_n is the actual attenuation value at time n.

The ADAM optimiser is employed to minimise the loss function [12]. To optimise the MLP model, the following hyper-parameters were tuned:

1. **Number of Hidden Layers and Neurons per Layer:** These parameters determine the complexity of the model. Too few layers or neurons result in an under-fitted model that cannot capture the problem's complexity, while too many lead to over-fitting.

2. **Activation Functions:** These functions transform input values of each layer into output values. Common choices include Sigmoid, Tanh (hyperbolic tangent), and ReLU (Rectified Linear Unit), as shown in Fig. 9 [8].

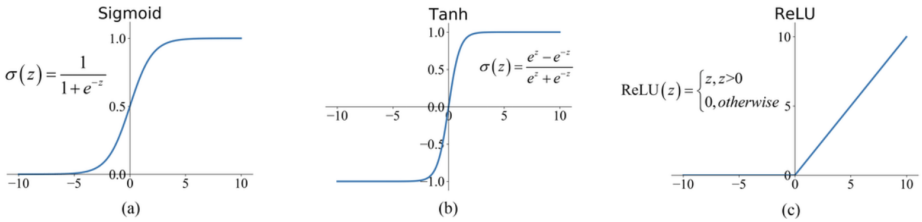


Figure 9: Common Activation Functions.

3. **Learning Rate:** A key parameter of the optimiser, the learning rate controls the step size for minimising the loss function. A higher learning rate accelerates convergence but risks overshooting the optimal solution, while a lower rate ensures precision but increases training time.

3.2.3 Convolutional Neural Network

Convolutional Neural Networks (CNNs) are a widely recognised deep learning architecture used in various fields. These networks comprise four key layer types: convolutional layers, pooling layers, activation layers, and fully connected layers (Fig. 10). In this study, CNN models were designed to focus on regression tasks, predicting signal attenuation in FSO and RF channels.

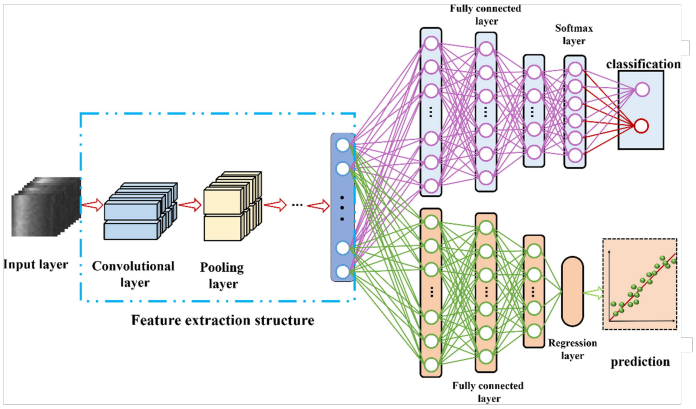


Figure 10: The basic structure for CNN. Adapted from [23]

The proposed CNN architecture consisted of a series of convolutional layers, each followed by a max pooling layer to reduce spatial dimensions and a batch normalisation layer to stabilise training. The output from the convolutional layers was flattened and passed through a fully connected layer, with dropout layers incorporated to mitigate over-fitting.

Mean Squared Error (MSE) was used as the loss function, and the Adam optimiser was employed, consistent with the MLP models. Key hyper-parameters were tuned to enhance model performance, including **Number of Convolutional Layers and Filters**, **Activation Functions** and **Learning Rate**.

3.3 Evaluation Metric

In this paper, each model is evaluated using the Root Mean Square Error (RMSE) and the R-squared value (R^2) below. A smaller RMSE and a larger R^2 indicate that the predicted attenuation closely aligns with the actual attenuation, reflecting better model performance.

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

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

where, N represents the total number of measurements,$\hat{y}_n$ is the predicted RF (or FSO) attenuation value at time n, y_n is the actual measured RF (or FSO) attenuation value at time n and $\bar{y}$ is the mean of the actual attenuation values over all n.

3.4 Workflow

The experimental work is structured into four main parts: data preparation, model creation, model comparison, and model evaluation. Separate models are developed for the FSO and RF channels, following an identical experimental process for both. The overall framework and workflow of the experiment are shown in Fig. 11.

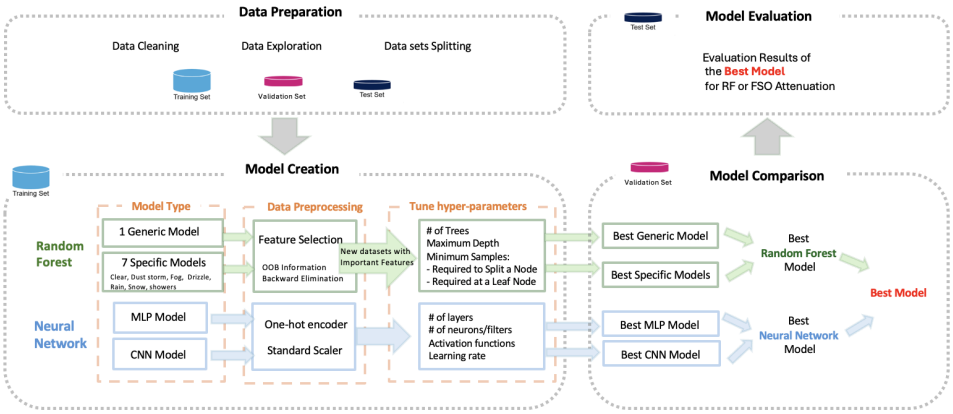


Figure 11: Presents the framework. Source: created by myself

4 Results

4.1 Data preparation

The synthetic empirical dataset contains a total of 91,379 measurements. For the Free Space Optical (FSO) system, FSO attenuation serves as the output variable, while for the Radio Frequency (RF) system, RF attenuation is used as the output variable. The RF system includes a total of 25 features, whereas the FSO system contains 24 features, as the frequency feature is exclusive to the RF system.

The dataset was split into a Training Set comprising 58,482 measurements (64% of the total) and a Validation Set comprising 14,621 measurements (16% of the total), together accounting for 80% of the dataset. The remaining 20% of the data, consisting of 18,276 measurements, was allocated to the Test Set.

Table 3: Number of Measurements by Weather Conditions in Training, Validation Sets

Weather Condition	**Training Set**	**Validation Set**	**Proportion**
SC=0, Clear	36,457	9,114	0.623
SC=3, Dust Storm	123	30	0.002
SC=4, Fog	298	75	0.005
SC=5, Drizzle	4,227	1,057	0.072
SC=6, Rain	16,011	4,003	0.274
SC=7, Snow	268	67	0.004
SC=8, Showers	1,098	275	0.019

To evaluate the performance of the specific models, the training and validation sets were further divided into subsets corresponding to the seven weather conditions. Table 3 presents the number of samples and their proportions for each weather condition.

During model creation and comparison, the training and validation sets were used to train model parameters and tune hyper-parameters. Final model performance was evaluated using the test set, which was not involved in the training process.

4.2 Random Forest Model Creation

4.2.1 Feature Selection

For the Random Forest models, both generic and specific models were constructed separately for the FSO and RF systems. Prior to training, feature selection was performed using the Algorithm 1.

Algorithm 1 Backward Elimination Algorithm

1: Let S be the initial set of N predictor variables.
2: Initialise an empty table R to store RMSE, R^2 values, and the ranking of features.
3: **while** S is not empty **do**
4: Train a Random Forest model with 100 trees using only the features in S.
5: Compute the OOB RMSE and R^2 values for the model.
6: Rank the features in S based on their OOB importance.
7: Remove the least important feature from S.
8: Record the removed feature, along with the RMSE and R^2 values, as a new row in R.
9: Output R, containing the feature rankings and the corresponding RMSE and R^2 values for each iteration.

Applying Algorithm 1, each model produced a table summarising the sequence of feature removal along with the corresponding OOB RMSE and R^2 values. Feature Importance Ranking Graphs were created by plotting RMSE and R^2 against the order of removed features for each model. Based on the observed trends, turning points were identified at positions where the RMSE curve increased most sharply, and the R^2 curve decreased most significantly. To highlight these critical points, green dotted lines were inserted into the graphs. Features to the right of the green line are considered important, while those to the left are deemed unimportant. Fig. 12, Fig. 13, and Fig. 14 show the feature importance ranking graphs for the generic and specific models of the FSO and RF systems, respectively. The important features selected for each FSO and RF model are summarised in Tab. 4.

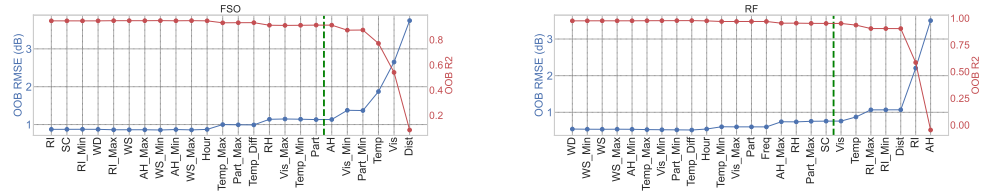
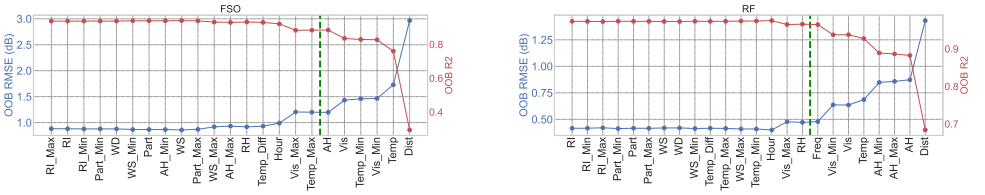
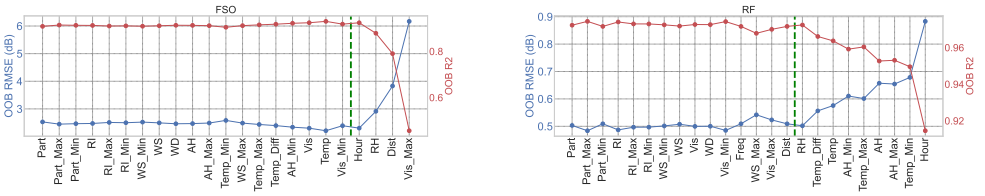


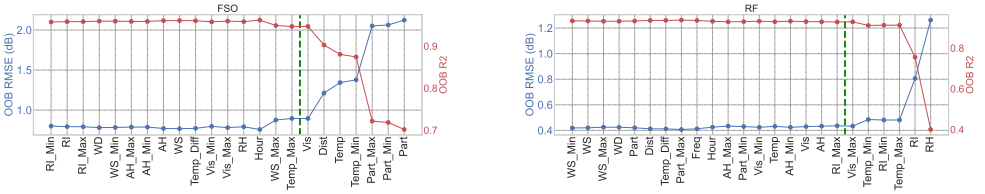
Figure 12: Feature Importance Rankings for Generic Models Across All Weather Conditions in FSO (left) and RF (right) Systems.



(a) Specific Models for Clear

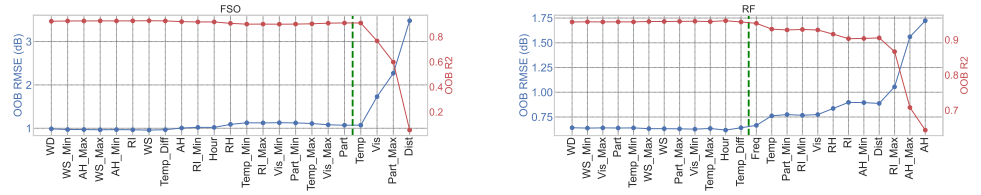


(b) Specific Models for Dust Storm

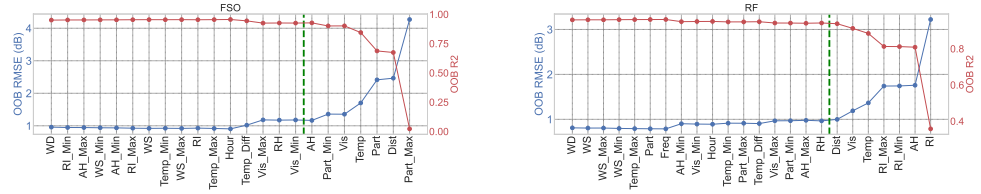


(c) Specific Models for Fog

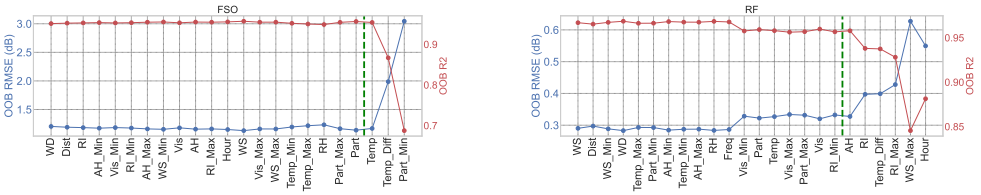
Figure 13: Feature Importance Rankings for Specific Models Across Different Conditions in FSO (left) and RF (right) Systems.



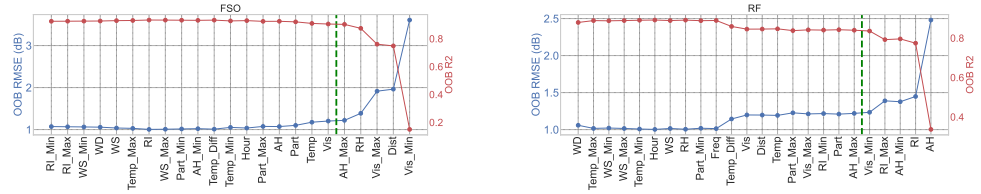
(a) Specific Models for Drizzle



(b) Specific Models for Rain



(c) Specific Models for Snow



(d) Specific Models for Showers

Figure 14: Feature Importance Rankings for Specific Models Across Different Conditions in FSO (left) and RF (right) Systems.

Table 4: Important Features for FSO and RF Models

(a) For FSO Models

Model	**Important Features**	**Count**
Generic	AH, Vis_Min, Part_Min, Temp, Vis, Dist	6
Specific for Clear	Hour, Vis_Max, Temp_Max, AH, Vis, Temp_Min, Vis_Min, Temp, Dist	9
Specific for Dust Storm	Hour, RH, Dist, Vis_Max	4
Specific for Fog	Vis, Dist, Temp, Temp_Min, Part_Max, Part_Min, Part	7
Specific for Drizzle	Temp, Vis, Part_Max, Dist	4
Specific for Rain	AH, Part_Min, Vis, Temp, Part, Dist, Part_Max	7
Specific for Snow	Temp, Temp_Diff, Part_Min	3
Specific for Showers	AH_Max, RH, Vis_Max, Dist, Vis_Min	5

(b) For RF Models

Model	**Important Features**	**Count**
Generic	Vis, Temp, RI_Max, RI_Min, Dist, RI, AH	7
Specific for Clear	Freq, Vis_Min, Vis, Temp, AH_Min, AH_Max, AH, Dist	8
Specific for Dust Storm	RH, Temp_Diff, Temp, AH_Min, Temp_Max, AH, AH_Max, Temp_Min, Hour	9
Specific for Fog	Vis_Max, Temp_Min, RI_Min, Temp_Max, RI, RH	6
Specific for Drizzle	Freq, Temp, Part_Min, RI_Min, Vis, RH, RI, AH_Min, Dist, RI_Max, AH_Max, AH	12
Specific for Rain	Dist, Vis, Temp, RI_Max, RI_Min, AH, RI	7
Specific for Snow	AH, RI, Temp_Diff, RI_Max, WS_Max, Hour	6
Specific for Showers	Vis_Min, RI_Max, AH_Min, RI, AH	5

From the figures and the summary tables of feature importance for each model, several expected trends are observed:

- Compared to RF, FSO is more susceptible to weather conditions. As a result, the performance of the FSO model is more significantly affected by feature removal. Additionally, the overall number of important features in the FSO model is smaller than that in the RF model.
- In weather conditions such as fog, drizzle, snow, and rain, the FSO model consistently identifies particle count (Part) as an important feature. This aligns with physical principles, as particles in the atmosphere can cause beam attenuation and scattering, leading to signal degradation.
- For the RF model, rain intensity (RI) is identified as an important feature under rain and snow conditions. This is consistent with the fact that the size and concentration of raindrops can significantly influence signal attenuation in RF systems.
- Visibility (Vis) and distance (Dist) are important features across both models under most weather conditions. Both features directly impact attenuation along the transmission path in FSO and RF channels.

Next, the original datasets were filtered into new sub-datasets based on the important features corresponding to each model, followed by training the respective random forest models.

4.2.2 Hyper-parameters Tuning

Following feature selection, GridSearchCV from the Scikit-learn library was employed to optimise the hyper-parameters of each Random Forest model using the updated training datasets. A 5-fold cross-validation approach was adopted.

During the hyper-parameter tuning process, one value was selected from a predefined set of options for each hyper-parameter. These hyper-parameter options, including the number of trees, maximum depth, minimum samples required to split a node, and minimum samples required at a leaf node, are detailed in Tab. 5. After hyper-parameter tuning, the optimal settings for each Random Forest model for both FSO and RF systems were determined, as shown in Tab. 6.

Table 5: Hyper-parameter Options for Random Forest Models

Hyper-parameters(Abbrv.)	**Listed Options**
Number of Trees(# Trees)	50, 100, 200
Max Depth(Max Depth)	10, 20
Min Samples Required to Split a Node(MS_Split)	2, 5
Min Samples Required at a Leaf Node(MS_Leaf)	1, 2

Table 6: Best Parameter Settings for Random Forest Models

(a) For FSO

Model	**# Trees**	**Max Depth**	**MS_Split**	**MS_Leaf**
Generic	200	20	2	1
Specific: Clear	200	20	2	1
Specific: Dust storm	50	20	2	1
Specific: Fog	50	20	2	1
Specific: Drizzle	200	20	2	1
Specific: Rain	200	20	2	1
Specific: Snow	100	10	2	1
Specific: Showers	50	20	2	1

(b) For RF

Model	**# Trees**	**Max Depth**	**MS_Split**	**MS_Leaf**
Generic	200	20	2	1
Specific: Clear	200	20	2	1
Specific: Dust Storm	100	20	2	1
Specific: Fog	200	20	2	1
Specific: Drizzle	200	20	2	1
Specific: Rain	200	20	2	1
Specific: Snow	50	10	2	1
Specific: Showers	50	20	5	1

4.3 Neural Network Model Creation

4.3.1 Data Preprocessing

Prior to training the neural network models, categorical variables, including SYNOP code, Hour, and Frequency, were one-hot encoded to transform them into a suitable numerical format. For continuous vari-

ables, which displayed right-skewed distributions, feature transformation was applied using the StandardScaler to standardise their scales.

4.3.2 Hyper-parameters Tuning

The neural network models were implemented in Keras using TensorFlow. All models employed the Adam optimiser and used MSE as the loss function, with RMSE and R^2 metrics to evaluate performance during training and validation. Training was conducted over 100 epochs with a batch size of 32 for Multi-Layer Perceptron (MLP) models, and 50 epochs with a batch size of 32 for Convolutional Neural Network (CNN) models.

The MLP models featured an input layer followed by a specified number of hidden layers, each containing a defined number of neurons and a specific activation function. Batch normalisation layers were incorporated after each hidden layer. The output layer consisted of a single neuron with a linear activation function, designed to predict attenuation values.

The CNN models included convolutional layers, each followed by max pooling and batch normalisation layers. The flattened outputs of the convolutional layers were passed through one fully connected layer. To reduce over-fitting, dropout layers were added, with a fixed dropout rate of 0.3. The specific hyper-parameter options for MLP and CNN models are summarised in Table 7.

Table 7: Hyper-parameter Options for MLP and CNN

(a) MLP

Hyper-parameters (Abbrv.)	**Listed Options**
Number of Hidden Layers(# Hidden Layers)	2, 3
Number of Neurons Per Layer(# Neurons)	32, 64, 128, 256
Activation Function (AF)	ReLU, Tanh, Sigmoid
Learning Rate (LR)	0.01, 0.001

(b) CNN

Hyper-parameters (Abbrv.)	**Listed Options**
Number of Convolutional Layers(# Conv Layers)	2, 3
Number of Filters Per Layer(# Filters)	128, 256
Activation Function (AF)	ReLU, Tanh, Sigmoid
Learning Rate (LR)	0.01, 0.001

A total of 480 hyper-parameter combinations were explored for MLP models and 72 combinations evaluated for CNN Models. Based on the results from hyper-parameter tuning, the best-performing MLP and CNN models for both the FSO and RF systems were selected and summarised in table Tab. 8.

Table 8: Best Model Parameter Settings

(a) for MLP

System	# Layers	# Neurons	AF	LR
FSO	3	(256, 128, 256)	sigmoid	0.001
RF	3	(256, 64, 256)	sigmoid	0.001

(b) for CNN

System	# Conv Layers	# Filters	AF	LR
FSO	3	(128, 128, 256)	sigmoid	0.001
RF	3	(128, 256, 256)	ReLU	0.001

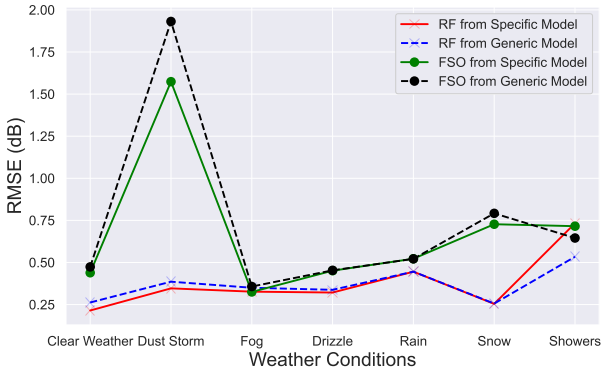
4.4 Model Comparison

4.4.1 Generic with Specific Random Forest Model

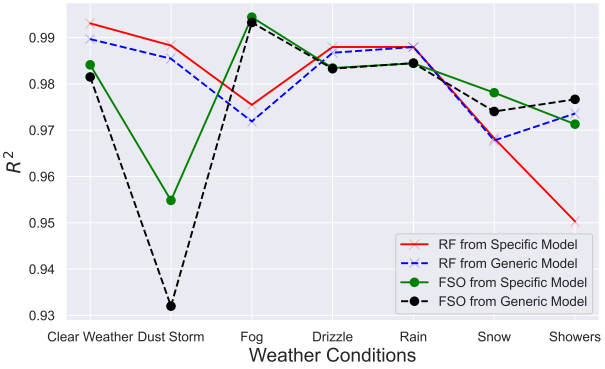
The performance of the Random Forest models, including both generic (GRF) and specific (SRF) models, was evaluated across different weather conditions using RMSE and R^2 as metrics.

From Fig. 15, it can be observed that the RF channel demonstrates significantly lower RMSE values compared to the FSO channel, highlighting its robustness and stability under varying weather conditions. In contrast, the FSO channel exhibits higher sensitivity to extreme conditions, such as dust storms, as indicated by sharp increases in RMSE and decreases in R^2 values.

Both RF and FSO channels experience performance degradation under dust storms, with SRF slightly outperforming GRF for both channels. For instance, in the FSO channel, SRF achieves an RMSE of 1.5737 and an R^2 of 0.9548, compared to GRF's RMSE of 1.6358 and R^2 of 0.9437. However, the limited proportion of dust storm data in the dataset poses challenges for both models in achieving optimal performance.



(a) RMSE



(b) R^2

Figure 15: Comparison of RMSE between Specific and Generic Models

Under showers, GRF consistently outperforms SRF for both RF and FSO channels. For the FSO channel, GRF achieves an RMSE of 0.6452 and an R^2 of 0.9767, outperforming SRF's RMSE of 0.7156 and R^2 of 0.9713. This demonstrates GRF's ability to leverage data from other weather conditions to improve performance when training data for a specific condition is scarce.

Apart from dust storms and showers, the performance of GRF and SRF is generally comparable for both RF and FSO channels. This suggests that the simpler GRF model is sufficient to achieve results similar to SRF across most weather conditions, making it a more practical and unified solution.

While SRF shows marginal advantages under specific conditions, such as dust storms, its significant under-performance in underrepresented

weather conditions, like showers, makes GRF the more robust and reliable choice. Therefore, GRF is selected as the best Random Forest model.

4.4.2 Random Forest with Neural Network Model

Tab. 9 presents the validation RMSE and R^2 metrics for the best-performing Random Forest (RF), Multi-Layer Perceptron (MLP), and Convolutional Neural Network (CNN) models for both FSO and RF systems. For the FSO system, the Random Forest model achieves the best performance with an RMSE of 0.5000 and an R^2 value of 0.9830, outperforming both the MLP and CNN models. The MLP model achieves an RMSE of 0.8678 and R^2 of 0.9384, while the CNN model has the highest RMSE of 0.9252 and the lowest R^2 value of 0.9309 among the three.

Table 9: Validation RMSE and R2 for Different Models

(a) For FSO

Model	**RMSE**	**R2**
Best Random Forest Model	0.5000	0.9830
Best MLP Model	0.8678	0.9384
Best CNN Model	0.9252	0.9309

(b) For RF

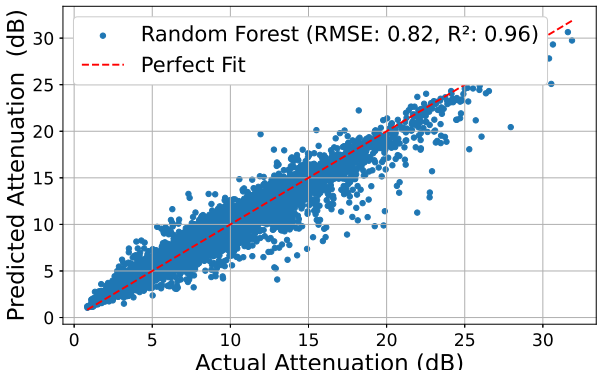
Model	**RMSE**	**R2**
Best Random Forest Model	0.3369	0.9906
Best MLP Model	0.491946	0.972022
Best CNN Model	0.586408	0.962392

Similarly, for the RF system, the Random Forest model again exhibits the best performance with an RMSE of 0.3369 and an R^2 value of 0.9906, significantly outperforming the MLP and CNN models. The MLP model achieves an RMSE of 0.4919 and R^2 of 0.9720, while the CNN model shows the lowest performance with an RMSE of 0.5864 and R^2 of 0.9624.

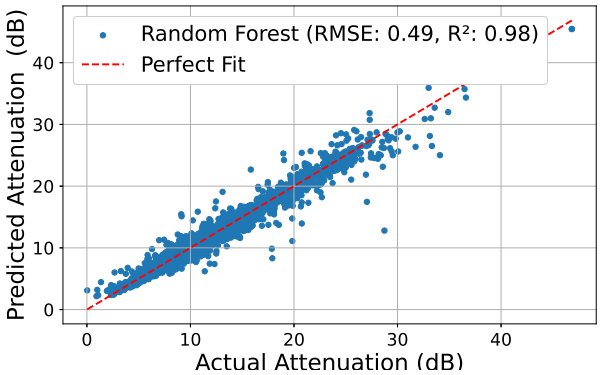
In summary, the generic Random Forest model is selected as the optimal attenuation prediction model for both the FSO and RF channels.

4.5 Model Evaluation

We evaluated our best model, the generic Random Forest model, on the test dataset. The model was first retrained using the combined training and validation datasets and subsequently evaluated on the test data. The performance of the models for predicting attenuation in the FSO and RF channels is illustrated in Fig. 16.



(a) For FSO



(b) For RF

Figure 16: Best Model Evaluation Results

For the FSO channel, the model achieved an RMSE of 0.82 dB and an R^2 value of 0.96. As shown in Figure Fig. 16a, the data points closely follow the ideal perfect-fit line (red dashed line). This alignment reflects the model's ability to effectively capture the underlying patterns in FSO attenuation with minimal error. However, slight deviations from the perfect-fit line, particularly at higher attenuation values, indicate that the model may face challenges in handling extreme cases. This is likely due to the high sensitivity of the FSO system to environmental factors.

For the RF channel, the model achieved an RMSE of 0.49 dB and an R^2 value of 0.98, demonstrating even better performance than the FSO model. As shown in Figure Fig. 16b, the data points show a near-perfect alignment with the ideal perfect-fit line. The lower RMSE value highlights the RF system's relative stability and robustness against environmental variations, making attenuation patterns easier for the model to predict with high accuracy.

5 Conclusion

This study developed predictive models for FSO and RF channel attenuation using two machine learning approaches: Random Forest and Neural Networks. Consistent methodologies were applied to both channels.

For the Random Forest models, two types of models were constructed for comparison: 1. Generic models, which treat weather conditions as a single categorical feature. 2. Specific models, which build separate models for each weather condition. To prevent over-fitting, feature selection was performed using a wrapper method with Backward Elimination based on Out-of-Bag (OOB) information. Hyper-parameter tuning was conducted for all Random Forest models using GridSearchCV, optimising parameters such as the number of trees, maximum depth, and minimum sample requirements. The best-performing generic Random Forest (GRF) and specific Random Forest (SRF) models were identified for both FSO and RF channels.

For Neural Network models, both feed-forward Multi-Layer Perceptrons (MLPs) and Convolutional Neural Networks (CNNs) were explored. Categorical features were processed using One-Hot Encoding, and continuous features were standardised to improve model training stability. Hyper-parameter optimisation was performed to determine the best architecture, including the number of units, hidden layers, activation functions, and learning rates. The optimal MLP and CNN models were selected for each channel.

Comparative evaluations on the validation set identified the GRF model as the best-performing Random Forest model for both FSO and RF channels. GRF demonstrated performance comparable to SRF across most weather conditions and consistently outperformed SRF in underrepresented scenarios like showers, offering a simpler, more practical, and unified solution. When compared with Neural Network models, the GRF model demonstrated superior validation results, establishing it as the best attenuation model under the current experimental settings.

The final evaluation on the test set showed strong performance for the best models. For the FSO channel, the GRF model achieved an RMSE of 0.82 dB and an R^2 value of 0.96, while for the RF channel, it achieved an RMSE of 0.49 dB and an R^2 value of 0.98

Although the study successfully identified the best attenuation mod-

els, there are areas for improvement:

1. Neural Network models could benefit from further exploration. Due to time constraints, only a limited number of architectures and hyper-parameter settings were tested. Future work could involve adjusting the learning rate, increasing the number of training epochs, or incorporating pre-trained models with fine-tuning.

2. For the FSO model, predictions tended to underestimate attenuation at higher values. Increasing the volume of data for higher attenuation levels or applying additional techniques and adjustments may help address this issue.

In the next phase of this research, we aim to build upon the established models by exploring the correlation between FSO and RF channels. This will provide deeper insights into how these channels respond to similar environmental conditions, paving the way for more robust hybrid communication systems.

Acknowledgements

I would like to thank Siu Wai Ho and my fellow group members for their invaluable support and collaboration throughout this research. Their guidance, feedback, and teamwork were instrumental in the successful completion of this study.

Additionally, I acknowledge the use of OpenAI's ChatGPT for providing writing assistance and aiding in the refinement of this manuscript.

A Appendices

Here is my code link for this paper:

GitHub link `https://github.com/ChrisLRY/Final-Project.git`.

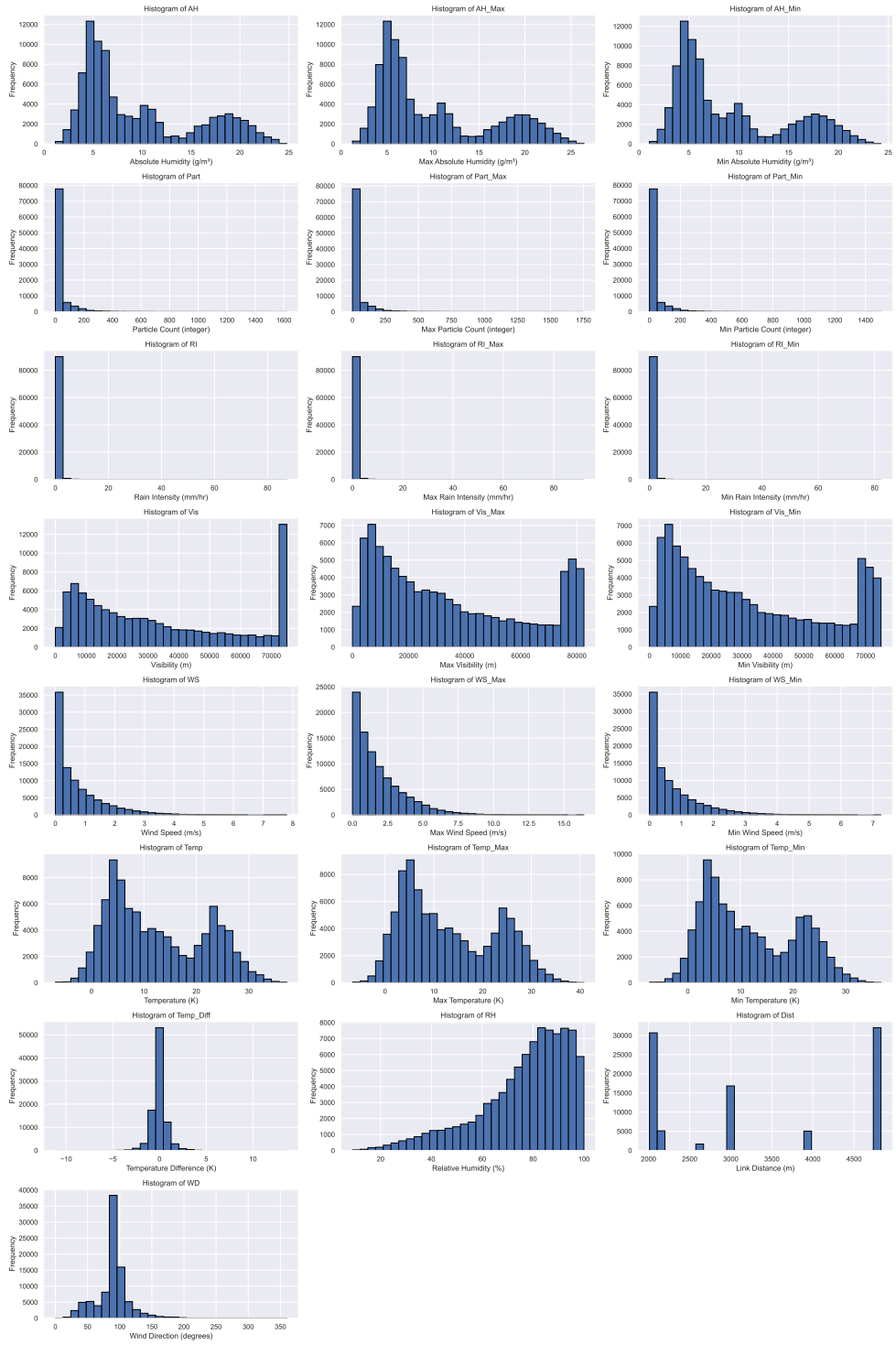
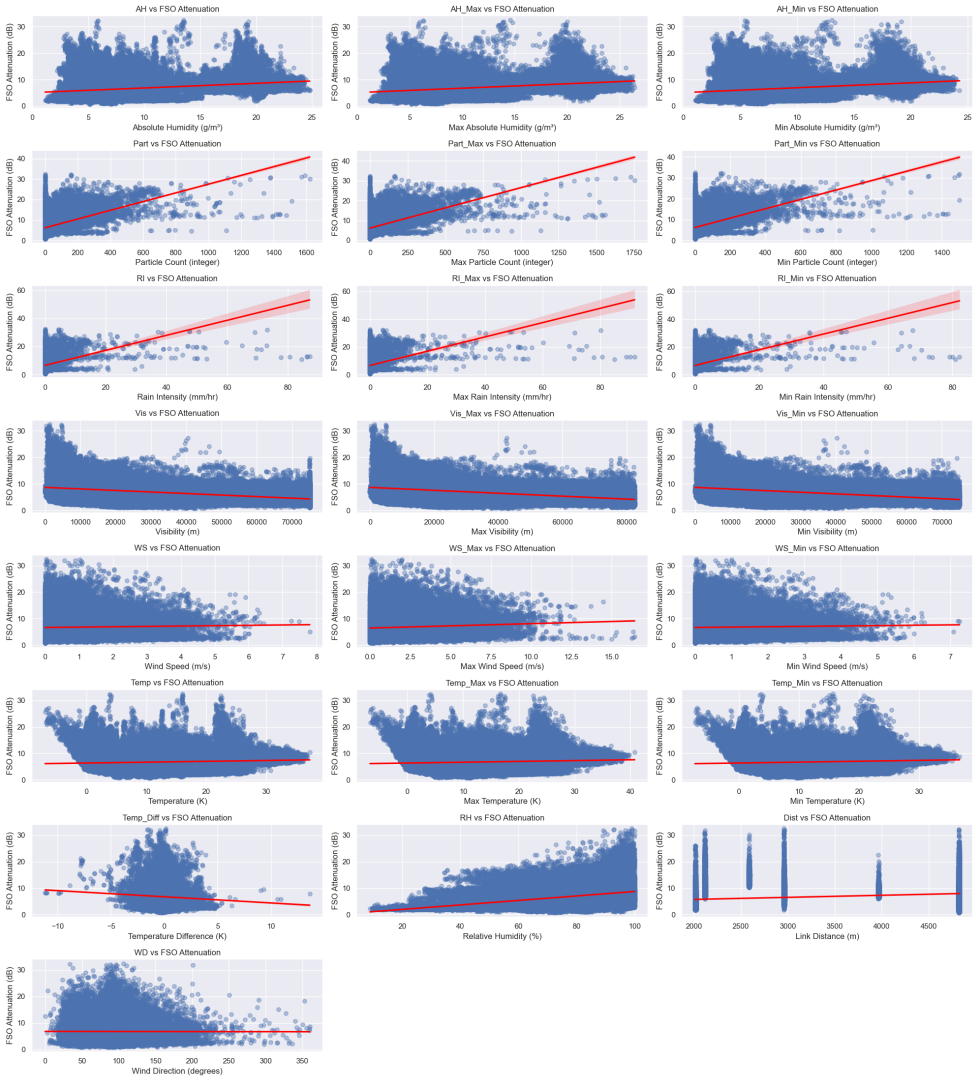
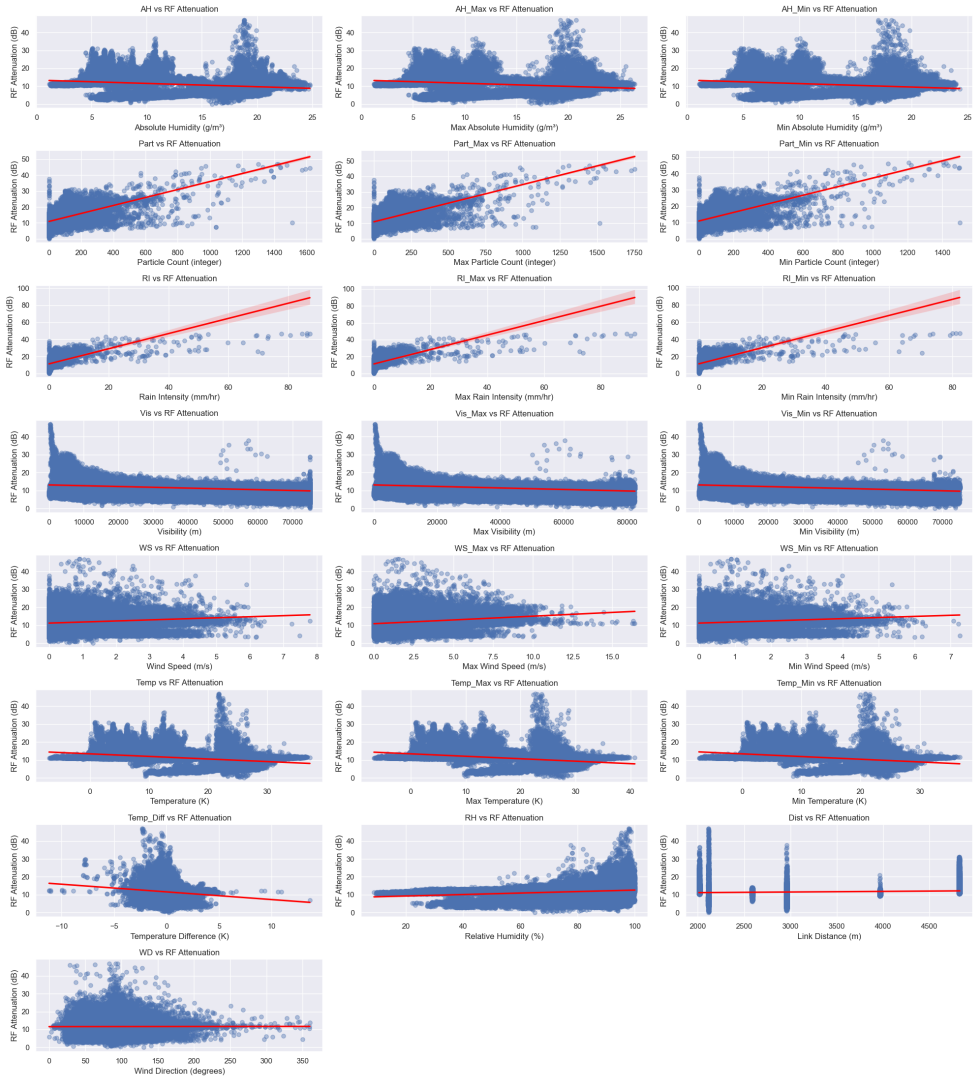


Figure 17: Histograms for Continuous Variables


AH vs FSO Attenuation
Absolute Humidity (g/m³)
AH_Max vs FSO Attenuation
Max Absolute Humidity (g/m³)
AH_Min vs FSO Attenuation
Min Absolute Humidity (g/m³)
Part vs FSO Attenuation
Particle Count (integer)
Part_Max vs FSO Attenuation
Max Particle Count (integer)
Part_Min vs FSO Attenuation
Min Particle Count (integer)
RI vs FSO Attenuation
Rain Intensity (mm/hr)
RI_Max vs FSO Attenuation
Max Rain Intensity (mm/hr)
RI_Min vs FSO Attenuation
Min Rain Intensity (mm/hr)
Vis vs FSO Attenuation
Visibility (m)
Vis_Max vs FSO Attenuation
Max Visibility (m)
Vis_Min vs FSO Attenuation
Min Visibility (m)
WS vs FSO Attenuation
Wind Speed (m/s)
WS_Max vs FSO Attenuation
Max Wind Speed (m/s)
WS_Min vs FSO Attenuation
Min Wind Speed (m/s)
Temp vs FSO Attenuation
Temperature (K)
Temp_Max vs FSO Attenuation
Max Temperature (K)
Temp_Min vs FSO Attenuation
Min Temperature (K)
Temp_Diff vs FSO Attenuation
Temperature Difference (K)
RH vs FSO Attenuation
Relative Humidity (%)
Dist vs FSO Attenuation
Link Distance (m)
WD vs FSO Attenuation
Wind Direction (degrees)
FSO Attenuation (dB)


AH vs RF Attenuation
Absolute Humidity (g/m³)
AH_Max vs RF Attenuation
Max Absolute Humidity (g/m³)
AH_Min vs RF Attenuation
Min Absolute Humidity (g/m³)
Part vs RF Attenuation
Particle Count (integer)
Part_Max vs RF Attenuation
Max Particle Count (integer)
Part_Min vs RF Attenuation
Min Particle Count (integer)
RI vs RF Attenuation
Rain Intensity (mm/hr)
RI_Max vs RF Attenuation
Max Rain Intensity (mm/hr)
RI_Min vs RF Attenuation
Min Rain Intensity (mm/hr)
Vis vs RF Attenuation
Visibility (m)
Vis_Max vs RF Attenuation
Max Visibility (m)
Vis_Min vs RF Attenuation
Min Visibility (m)
WS vs RF Attenuation
Wind Speed (m/s)
WS_Max vs RF Attenuation
Max Wind Speed (m/s)
WS_Min vs RF Attenuation
Min Wind Speed (m/s)
Temp vs RF Attenuation
Temperature (K)
Temp_Max vs RF Attenuation
Max Temperature (K)
Temp_Min vs RF Attenuation
Min Temperature (K)
Temp_Diff vs RF Attenuation
Temperature Difference (K)
RH vs RF Attenuation
Relative Humidity (%)
Dist vs RF Attenuation
Link Distance (m)
WD vs RF Attenuation
Wind Direction (degrees)
RF Attenuation (dB)

References

[1] M. M. Abadi and Z. Ghassemlooy. A hybrid free-space optical (fso)/radio frequency (rf) antenna for satellite applications. In *2022 13th International Symposium on Communication Systems, Networks and Digital Signal Processing (CSNDSP)*, pages 185–190, 2022. 2

[2] A. S. Alatawi, A. A. Youssef, M. Abaza, et al. The effects of atmospheric turbulence on optical wireless communication in neom smart city. *MDPI Photonics*, 9:262, 2022. 4

[3] M. Z. Chowdhury, M. K. Hasan, M. Shahjalal, M. T. Hossan, and Y. M. Jang. Optical wireless hybrid networks: Trends, opportunities, challenges, and research directions. *IEEE Communications Surveys & Tutorials*, 22:930–966, 2020. 2

[4] C. Coleman. Basic concepts. In *An Introduction to Radio Frequency Engineering*, pages 1–27. Cambridge University Press, Cambridge, 2004. 4

[5] Matthias Feurer and Frank Hutter. Hyperparameter optimization. In *Automated Machine Learning: Methods, Systems, Challenges*, pages 3–33. Springer, 2019. 11

[6] Tin Kam Ho. Random decision forests. In *Proceedings of the 3rd International Conference on Document Analysis and Recognition*, pages 278–282, Montreal, QC, 1995. 9

[7] TK Ho. The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(8):832–844, 1998. 9

[8] Shruti Jadon. Introduction to different activation functions for deep learning, 2018. 14

[9] A. G. Kanatas and A. D. Panagopoulos. Radio wave propagation and channel modeling for earth–space systems. *System*, 30(2):76–87, 2012. 4

[10] A. G. Karegowda, M. A. Jayaram, and A. S. Manjunath. Feature subset selection problem using wrapper approach in supervised learning. *International Journal of Computer Applications*, 1(7):13–17, 2010. 10

[11] H. Khalid, S. M. Sajid, H. E. Nistazakis, and M. Ijaz. Survey on limitations, applications and challenges for machine learning aided hybrid fso/rf systems under fog and smog influence. *Journal of Modern Optics*, 71(4–6):101–125, 2024. 4

[12] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv.org*, 2017. arXiv preprint arXiv:1412.6980. 14

[13] R. Kohavi and G. H. John. Wrappers for feature subset selection. *Artificial Intelligence*, 97(1):273–324, 1997. 10

[14] P. Krishnan. Performance analysis of hybrid rf/fso system using bpsk-sim and dpsk-sim over gamma-gamma turbulence channel with pointing errors for smart city applications. *IEEE Access*, 6:75025–75032, 2018. 2

[15] S Kumar, P Kaur, and A Gosain. A comprehensive survey on ensemble methods. In *Proceedings of the IEEE 7th International Conference for Convergence in Technology (I2CT)*, pages 1–7, Mumbai, India, 2022. 9

[16] W. Li. Interpretation and classification of p-series recommendations in itu-r. *International Journal of Communications, Network and System Sciences*, 9(5):117, 2016. 2

[17] Jie Liang, Jiahao Chen, Xueqin Zhang, Yue Zhou, and JiaJun Lin. Anomaly detection based on singular heat coding and convolutional neural network. *Journal of Tsinghua University (Natural Science Edition)*, 25:1–7, 2018. 12

[18] A. Mahdy and J. S. Deogun. Wireless optical communications: A survey. In *2004 IEEE Wireless Communications and Networking Conference (IEEE Cat. No. 04TH8733)*, volume 4, pages 2399–2404. IEEE, March 2004. 2

[19] MathWorks. Mathworks function treebagger. Accessed: 2022-11-05. 11, 13

[20] ID Mienye and Y Sun. A survey of ensemble learning: Concepts, algorithms, applications, and prospects. *IEEE Access*, 10:99129–99149, 2022. 9

[21] S. Noreen, F. Giannetti, and V. Lottici. Earth and martian sand and dust storms: A comprehensive review of attenuation modelling and measurements. *IEEE Access*, 12:878–922, 2024. 4

[22] R. Umar, S. S. Sulan, A. W. Azlan, Z. A. Ibrahim, W. Z. A. W. Mokhtar, and N. H. Sabri. Radio frequency interference: The study of rain effect on radio signal attenuation. *Malaysian Journal of Analytical Sciences*, 19(5):1093–1098, 2015. 4

[23] Hongfei Zhu, Lianhe Yang, Jiyue Gao, Mei Gao, and Zhongzhi Han. Quantitative detection of aflatoxin b1 by subpixel cnn regression. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 268:120633, 2022. 15