(Received 1 August 2025, Accepted 22 October 2025, Scheduled 14 December 2025)

Rainfall DSD Modelling Using Supervised Learning Techniques for Rain Attenuation Prediction in South Africa

Tsietsi C. Ramatladi and Akintunde A. Alonge*

Department of Electrical Engineering Technology, University of Johannesburg, Johannesburg, South Africa

ABSTRACT: Reliable modelling of rainfall-induced attenuation is crucial for designing and operating high-frequency communication systems, particularly those operating above 10 GHz, in regions with severe rainfall conditions such as subtropical climates. This study offers a comparison of supervised machine learning (ML) models — k-nearest neighbours (KNN), decision trees (DT), and random forests (RF) — against traditional statistical methods such as the lognormal and gamma distributions for estimating raindrop size distribution (DSD) and specific attenuation. Rainfall measurements taken between 2018 and 2019 were obtained from a 1-minute disdrometer at the measurement location in Durban, South Africa (29.8651°S, 30.9734°E). The investigated models were then tested across four different rainfall regimes processed from the dataset: drizzle, widespread rain, shower and thunderstorm. An adaptive tuning method for selecting the best k-value in KNN was introduced to enhance prediction accuracy across various rainfall intensities. Model performances are evaluated using metrics such as root mean square error (RMSE), mean absolute error (MAE), and the coefficient of determination (R^2). The results show that KNN outperforms RF and DT, providing the highest accuracy and lowest prediction errors across all rainfall regimes. The confidence interval (CI) analysis confirms that KNN delivers more precise and stable estimates, while RF and DT exhibit greater variability and uncertainty in performance. Additionally, specific attenuation estimations from these ML models are compared at different rain rates with the ITU-R P.838-8 estimations for frequencies up to 100 GHz. These findings highlight the superiority of data-driven model, particularly the adaptive KNN, in capturing complex rainfall microstructures and improving attenuation predictions. This has direct implications for planning and deploying rain-resilient wireless networks in variable climatic regions.

1. INTRODUCTION

Wireless communication systems have become fundamental to everyday human life and are crucial for essential sectors such as finance, healthcare, and transportation. These systems employ wireless signals in the form of travelling electromagnetic waves to transmit and receive information through the propagation medium. However, when these signals are disrupted — through attenuation, interference, or other propagation impairments — the reliability of connectivity is compromised, which can lead to severe interruptions in these critical services. The past decade has seen an increasing global dependence on communication technologies. This trend has driven the demand for expanded network capacity, particularly in high-frequency systems operating within the microwave and millimetre-wave bands, between 10 GHz and 100 GHz. However, this shift introduces new challenges, notably signal propagation impairments caused by atmospheric conditions such as rainfall, humidity, fog, dust and turbulence, which lead to attenuation. Accurate modelling of these impairments, along with the application of suitable mitigation techniques such as adaptive power control, site diversity, and fade margin design, is critical to ensuring system reliability and performance. In rain-prone regions within the South African territory, studies have shown that rainfall-induced attenuation can significantly degrade signal quality, resulting in reduced system availability and capacity [1, 2].

Rainfall is the result of water droplets formed in clouds which return to the Earth surface as precipitation. Thus, the intensity, duration, and the distribution of rain drops vary both spatially and temporally. Rainfall-induced attenuation is primarily due to the scattering and absorption of electromagnetic waves by raindrops — a process governed by the DSD and its interaction with signal frequency along the propagation path [3]. DSD modelling serves as the cornerstone for estimating rainfall-specific attenuation, which quantifies the power loss per unit distance due to rain [4]. Traditional statistical approaches have been used to model local DSDs by fitting empirical data to parametric distributions such as the lognormal and gamma DSD models [4–9]. These methods provide simplicity and physical interpretability; however, they often struggle to capture the full variability of DSD characteristics, particularly in cases where the relationships between drop sizes and other rainfall microparameters are highly non-linear and embedded within large complex datasets. Traditional methods often rely on manual feature engineering, distribution fitting, or iterative statistical techniques to handle large volumes of data, which can be computationally expensive or even impractical. This highlights the need for more powerful, efficient, and scalable approaches such as machine learning (ML). By utilizing scalable algorithms and computational frameworks, ML can reveal hidden structures within data, deliver robust predictions, and adapt to various do-

* Corresponding author: Akintunde Ayodeji Alonge (aalonge@uj.ac.za).

mains characterized by complex, high-dimensional variability [10]. ML models can effectively capture intricate dependencies without requiring explicit functional assumptions while learning patterns directly from the data.

To overcome the deficiencies of statistical DSD models, emerging approaches related to machine learning techniques are robust alternatives for data processing and modelling. These models have been effectively employed in various fields, especially in environmental monitoring, hydrometeorological modelling, radio system design and radar meteorology [11–13]. Numerous studies have investigated statistical methods for modelling DSD and estimating specific attenuation, but the application of machine learning (ML) models in this domain remains limited and underexplored. Unlike traditional statistical approaches, ML models can factor in the complex non-linear relationships within datasets and generalize patterns across diverse features and labels, without relying on parametric assumptions [14–16]. Likewise, supervised ML algorithms such as k-nearest neighbours (KNN), decision trees (DT), and random forests (RF) have shown strong potential for producing accurate predictions with high computational efficiency.

This current study seeks to close the gap by introducing three supervised machine learning models — KNN, DT, and RF — while comparing their performances against traditional lognormal and gamma models for estimating DSD and specific attenuation prediction. The study area is Durban, a coastal city in South Africa located at the coordinates, 29.8651°S, 30.9734°E. The city experiences a humid subtropical climate, which is classified as Cfa under the Köppen climate classification system [http://stepsatest.csir.co.za]. The climate is modulated by the Agulhas current within the eastern Indian Ocean, with annual rainfall exceeding 1000 mm [2]. Thus, the city expanse is characterized by hot, humid summers and mild winters, with relatively consistent rainfall throughout the year. Due to its geographical location, it ranks among the wettest urban areas in South Africa, showing notable variability in rainfall patterns across the seasons. This study utilizes disdrometer dataset collected in Durban from 2018 to 2019 to evaluate models across four distinct rainfall regimes: drizzle, widespread rain, showers, and thunderstorms. The results are evaluated based on metric measures and their alignment with empirically derived attenuation values from measured data. This work contributes to the growing evidence that data-driven methods can outperform traditional statistical models in capturing real-world rainfall microstructures, with direct implications for the design of rain-resilient wireless networks in subtropical climates.

2. RELATED WORKS

Marshall and Palmer (1948), in their earliest works, defined DSD as an important tool for determining the size variation and concentration of raindrops in a specified volume of air during a rainfall event. It is an essential parameter for understanding the influence of rainfall on electromagnetic wave propagation, particularly in terrestrial (microwave transmission) and satellite communication systems [13]. Several traditional DSD models are derived from popular statistical distributions, such as the negative exponential, lognormal, and gamma distributions, among others, which are widely used. Afullo [17] and Owolawi [7], in their study, expressed the general form of DSD as:

$$N\left(D\right) = N_c \times f\left(D\right) \tag{1}$$

where $N\left(D\right)$ denotes the density of raindrops per unit volume for each size of diameter, usually represented in ($\mathrm{m}^{-3}\mathrm{mm}^{-1}$). The constant N_c is a normalisation factor or scaling constant dependent on rainfall intensity and environmental parameters.

(1) describes the variation of raindrop accumulation within diametric clusters in a unit volume of air. Various empirical and theoretical models employ different mathematical functions to represent the drop size probability density function, $f(D)$, accurately. To this end, several statistical models have been developed to describe DSD, each with different assumptions and fitting techniques. The negative exponential model, introduced by Marshall and Palmer [18], is recognised as one of the earliest and most widely used DSD models with the formulation given as:

$$N\left(D\right) = N_0 e^{-\Lambda D} \tag{2}$$

N_0 represents the intercept parameter, expressed in ($\mathrm{mm}^{-1}\mathrm{m}^{-3}$), which defines the instant concentration (m^{-3}) of raindrops at a given diameter. The slope parameter, Λ, measured in mm^{-1}, characterizes the rate of decrease in raindrop concentration as the drop size increases. Lastly, the parameter D denotes the raindrop diameter in millimetres (mm), which is a key variable in determining the distribution and behaviour of raindrops within a given rainfall event.

Other relevant models for characterizing rainfall DSD have been developed by several authors [5, 6]. One such model is the gamma distribution, which offers improved flexibility over simpler models like the exponential distribution. The model introduces an additional shape parameter μ, to enable the accurate representation of a broader range of rainfall conditions. The gamma DSD is mathematically expressed as [5]:

$$N\left(D\right) = N_0 D^{\mu} e^{-\Lambda D} \tag{3}$$

wherein the empirical regression power-law equations are given as thus,

$$N_o = aR^b \tag{3a}$$

$$\Lambda = a_\Lambda R^{b_\Lambda} \tag{3b}$$

The gamma DSD model is confirmed to perform well under convective rainfall conditions marked by high rain rates in tropical and subtropical regions [8, 9, 19].

The lognormal DSD model represents another statistical model; it assumes that raindrop diameters follow a lognormally distributed probability function. Ajayi and Olsen [6] employed this DSD model in a tropical region study as defined:

$$N\left(D\right) = \frac{N_t}{\sigma D\sqrt{2\pi}} \exp\left[-\frac{\left(\ln D - \mu\right)^2}{2\sigma^2}\right] \tag{4}$$

The parameter N_t represents the overall sum of drops across all diameter classes. The parameters μ, σ, and N_t are typically estimated from empirical DSD measurements using the method of moments (MoM) [2, 4, 11]. The result is a set of empirical power-law relationships that link the DSD parameters to rain-

fall rate R as follows:

$$N_t = aR^b \tag{4a}$$
$$\mu = A_\mu + B_\mu \ln R \tag{4b}$$
$$\sigma^2 = A_\sigma + B_\sigma \ln R \tag{4c}$$

The lognormal distribution has also demonstrated superior performance in modelling the natural variability of DSD, particularly in subtropical and tropical regions characterised by intense and highly variable rainfall conditions [6, 8, 9, 19, 20].

3. INTRODUCTION TO MACHINE LEARNING APPROACHES FOR DSD MODELLING

Traditional DSD models depend on fixed equations — with underlying statistical characteristics — which often restrict their ability to adapt to complex and dynamic meteorological conditions. Machine learning (ML), however, offers an alternative approach by learning patterns directly from rainfall data to reveal the complex relationships among feature variables and labels [16].

3.1. The k-Nearest Neighbours (KNN) for DSD Estimation

KNN is a versatile learning method based on instances, commonly utilized for classification and regression tasks [21]. KNN, as a lazy learning method, does not build a global model during training. Instead, it keeps the complete training dataset and performs calculations solely when generating a prediction. When a new input, x, is presented, the algorithm locates the closest training samples using a specified distance measure and derives the prediction from their associated outputs. In classification problems, KNN assigns a class to a new input by majority voting among its k nearest neighbours. Given a training set:

$$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n} \tag{5}$$

where each $x_i \in \mathbb{R}^d$ and $y_i \in \{1, 2, 3, \ldots, \mathcal{C}\}$, the predicted class $\hat{y}$ is defined as:

$$\hat{y} = \underset{c \in \{1,\ldots,\mathcal{C}\}}{\arg\max} \sum_{i \in \mathcal{N}_k(x)} \mathbb{I}(y_i = c) \tag{6a}$$

where $\mathcal{N}_k(x)$ is the set of the k nearest neighbours of x, and $\mathbb{I}(\ldots)$ is the indicator function.

In regression, KNN determines the output for a specified input, x, by averaging the output values of its k nearest neighbors in the training set. The fundamental idea behind KNN is based on the premise that data points sharing similar input features often exhibit related output values, ensuring the proximity of observations in the feature space is essential for accurate predictions [21]. For regression, the prediction is the mean of the output values of the neighbours:

$$\hat{y} = \frac{1}{k} \sum_{i \in \mathcal{N}_k(x)}^{k} y_i \tag{6b}$$

The accuracy of KNN is heavily influenced by the distance metric employed. Commonly used metrics include Euclidean distance [14, 21, 22],

$$d(x, x_i) = \sqrt{\sum_{j=1}^{d} (x_j - x_{i,j})^2} \tag{7a}$$

the Manhattan distance,

$$d(x, x_i) = \sum_{j=1}^{d} |x_j - x_{i,j}| \tag{7b}$$

and the general form is known as the Minkowski distance,

$$d(x, x_i) = \left(\sum_{j=1}^{d} |x_j - x_{i,j}|^p \right)^{1/p} \tag{7c}$$

Presume that the sequence of the norm is defined by p.

KNN performs well with both numerical and categorical data, especially when the data is low-dimensional and follows a relatively simple distribution [22]. However, its effectiveness can be impacted by differences in feature scales, which makes data preprocessing, such as normalization or standardization, essential. The key strengths of this algorithm lie in its straightforward implementation, the absence of a training phase, and its ability to handle multi-class classification tasks naturally. However, it requires substantial computational time during prediction, especially with large or high-dimensional datasets. This issue is widely referred to as the "curse of dimensionality", highlighting the decline in performance that occurs with an increasing number of input features [23].

3.2. Decision Trees (DT) for DSD Estimation

The DT model employs feature space splitting with axis-aligned divisions, thereby creating separate regions for which predictions are made [15, 24]. It creates a decision tree where each non-leaf node partitions the data using a feature-value comparison. This method applies to both classification and regression tasks, offering an intuitive, rule-based approach to decision-making.

In classification problems, widely used algorithms such as ID3 (Iterative Dichotomiser 3) and CART (Classification and Regression Trees) rely on impurity measures to guide the splitting process. ID3 is a decision tree algorithm that employs the Top-Down Induction of Decision Trees (TDIDT) method, which is designed to efficiently produce detailed decision trees from datasets containing numerous instance attributes. It relies on an entropy-based method to evaluate splits, selecting the attribute that offers the highest information gain at each decision point [24]. In contrast, CART employs Gini impurity as its splitting criterion [15]. Gini impurity is defined as:

$$G(S) = 1 - \sum_{c=1}^{C} p(c)^2 \tag{8}$$

While the entropy measures the amount of impurity or disorder in dataset using,

$$H(S) = -\sum_{c=1}^{C} p(c) \log 2p(c) \tag{9}$$

where $p(c)$ represents the fraction of samples in the node that belong to class (c) in node S. The split at a given feature is chosen to maximize Information Gain (IG):

$$IG(S, A) = H(S) - \sum_{v \in Values(A)} \frac{\mid S_v \mid}{\mid S \mid} H(S_v) \tag{10}$$

where S_v is the subset of samples in S with feature $A = v$.

For regression tasks, decision trees use variance reduction as a splitting criterion. The variance of a node S is calculated as:

$$Var(S) = \frac{1}{|S|} \sum_{i \in S} (y_i - \bar{y})^2 \tag{11}$$

where $\bar{y}$ refers to the average predicted value for observations in node S. A split is selected to minimize the weighted average variance of the resulting child nodes.

To avoid overfitting, decision trees apply pruning techniques. Pre-pruning prevents over-complex trees by imposing limits such as how deep the tree can grow or how many samples a node must contain to be split. In contrast, post-pruning techniques, such as cost-complexity pruning, trim back branches after the entire tree has grown, removing those with minimal predictive contribution [15].

3.3. Random Forest (RF) for DSD Estimation

Random Forest (RF) is an ensemble technique that enhances prediction accuracy and reduces overfitting by building multiple decision trees and combining their outputs [25]. At each node where a split decision is made, the algorithm considers only a randomly chosen subset of the available features, selecting a limited number of features from the complete feature set. Mathematically, for a forest containing B trees, each tree T_b is trained on a unique bootstrap sample S_b drawn from the complete dataset $\mathcal{D}$. The predicted class in classification is determined by aggregating the votes from all trees and selecting the majority class [25]:

$$\hat{y} = \arg\max_{c} \sum_{b=1}^{B} \mathrm{I}(T_b(x) = c) \tag{12}$$

where $I(..)$ is the indicator function. In regression tasks, the output represents the mean of the predictions from individual trees,

$$\hat{y} = \frac{1}{B} \sum_{b=1}^{B} T_b(x) \tag{13}$$

A significant advantage of RF is its capability to estimate the feature importance. This is generally calculated by considering the overall decrease in impurity attributed to each feature throughout all the trees. For a given feature j, its importance is calculated as [26],

$$importance(j) = \sum_{t \in splits\ on\ j} \frac{N_t}{N} \Delta i_t \tag{14}$$

where Δi_t represents the impurity decrease at node t, N_t refers to the count of samples reaching a particular node, and N is the total sample size of the dataset. RF excels with high-dimensional datasets and can handle both numerical and categorical data variables. However, it is less interpretable than a single decision tree and can be computationally demanding, especially when the number of trees or data size is large [27].

4. METHODOLOGY

The following outlines the procedure for training ML models to estimate DSD across different drop diameter bins in four rainfall regimes. The predicted values are used in Mie scattering calculations to determine the extinction cross-sections (ECS), which are then employed to estimate rainfall-induced attenuation.

4.1. Data Acquisition and Preprocessing

This research primarily used data collected from an impact disdrometer, specifically the RD-80 Joss-Waldvogel impact disdrometer, developed by Disdromet Ltd in Switzerland. The device was installed at the University of KwaZulu-Natal in Durban, which is known for its humid subtropical climate and distinct seasonal variations rainfall. The data collection process lasted for two years, spanning from January 2018 to December 2019. The equipment consists of two main components, the outdoor sensing unit which detects the raindrop impacts via a sensor and generates corresponding electrical signals — and the indoor processing unit which processes the received signal by categorizing them into specified raindrop sizes for data storage. Measurement of rainfall microphysical properties such as rain rate, rainfall DSD, and accumulated rainfall are done in real-time. The sensing area, covering 50 cm^2, captures raindrops at 1-minute interval and classify them into 20 separate diameter bins ranging from 0.359 to 5.373 mm. The precision of the instrument ensures measurement accuracy within 5% of actual values. The disdrometer computes rainfall rate by using the equation [28]:

$$R = \frac{6\pi \times 10^{-4}}{A \times T} \sum_{i=1}^{20} n_i \times D_i^3 \ [\mathrm{mm/h}] \tag{15}$$

where A is the area from which the sensor collects data, T the duration for sampling, D_i the drop diameter for the size class i, and n_i the quantity of drops documented in class i. The drop number density $N(D_i)$, which defines the drop size distribution (DSD) for diameter classes, D_i, is calculated as described [28],

$$N(D_i) = \frac{n_i}{A \times T \times v(D_i) \times \Delta D_i} \ [\mathrm{m}^{-3}\mathrm{mm}^{-1}] \tag{16}$$

where $v(D_i)$ represents the drop velocity, and ΔD_i indicates the diameter width measured in mm.

The obtained rainfall data was further processed into four distinct rainfall regimes based on the recorded maximum rain rate per DSD spectra: drizzle (for rains less than 5 mm/h), widespread (for rains between 5 mm/h and 10 mm/h), showers (for rain between 10 mm/h and 40 mm/h), and thunderstorms (for rains greater than 40 mm/h) [29]. Samples with a rain rate below 0.1 mm/h and fewer than 10 total drops were excluded to minimize the impact of dead time errors [30]. The disdrometer dataset was carefully divided into training, optimal k tuning, and validation subsets wherein a 56 : 16 : 28 ratio was applied to ensure thorough and reliable model evaluation.

4.2. Optimising k in KNN for a Dynamic Approach to DSD Modelling

The KNN algorithm works by finding and averaging the values of the k-nearest data points in the training dataset. Its performance heavily depends on the hyperparameter k, which sets the number of neighbouring points used for prediction. It is implemented with the *Kneighborsregressor* module from Scikit-Learn machine learning library. The proposed KNN model uses the weighted approach introduced in [14], assigning importance to neighbouring points based on their distance, as shown in the following equation:

$$N_D(X) = \frac{\sum_{i=1}^{k} w_i N_D(X_i)}{\sum_{i=1}^{k} w_i} \tag{17}$$

where $N_D(X)$ denotes the values corresponding to the k closest data points from the training set. This approach uses a weight function, w_i, that is dependent on the distance $d(X, X_i)$ between the test sample X and neighbour X_i:

$$w_i = \frac{1}{d(X, X_i)} \tag{18}$$

The weighted method is particularly effective when the training data is unevenly distributed. The hyperparameter k must be predefined before the training and predictions to ensure optimal performance. This study introduces an adaptive method for choosing k, allowing the model to automatically select the best k-value to improve prediction accuracy and reduce errors.

4.3. Adaptive Tuning for Optimal k Parameter

To enhance the adaptability of the proposed KNN model, an adaptive tuning strategy was employed to identify the optimal number of neighbours (k). This process involves using the k-tuning subset reserved for each rainfall regime to train the model to dynamically select the most appropriate k-value by analysing the statistical properties of each individual sample. For this, each feature vector $x_i \in \mathbb{R}^d$ represents a data point situated in a d-dimensional space, where d corresponds to the number of rainfall-microparameter variables such as drop diameter, rainfall rate, and related characteristics. By treating each input (feature) as a vector in $\mathbb{R}^d$, the KNN algorithm computes distances to its neighbours and predicts outcomes (DSD) based on their proximity.

The tuning process for determining the optimal k value is performed using a k-fold cross-validation method. In this approach, the tuning subset is divided into folds. The model is trained on, $k_{cv} - 1$, folds for each candidate value of k, which ranges from 1 to 100, and validated on the remaining data fold. This procedure is carried out for each fold, with the cross-validation error calculated for every averaged candidate k. The optimal value, k^*, is then chosen using the following criterion given in [14],

$$k^* = \arg\max_{k \in \mathcal{K}} CVError(k) \tag{19}$$

where $\mathcal{K} = \{1, 2, 3, 4, \ldots, 100\}$ represents the set of preselected candidate k values, and $CVError(k)$ denotes the average validation error across all folds for that k. The aim is to reduce expected prediction error and enhance model generalization through empirical risk minimization framework.

During the tuning process, the mean squared error (MSE) serves as the performance metric and is calculated as follows [21]:

$$\text{MSE}(k) = \frac{1}{m} \sum_{j=1}^{m} \left(y_j - \hat{y}_j^{(k)} \right)^2 \tag{20}$$

where m denotes the quantity of validation samples, and $\hat{y}_j^{(k)}$ represents the anticipated value for the j^{th} sample using k-nearest neighbours. This approach is often referred to as the "elbow method" because it visually highlights the point where increasing the value of k no longer results in significant improvements in error reduction.

Once the optimal k value is determined for each sample within the k-tuning dataset across different rainfall regimes, the results are compiled into a reference dataset that links each specific rainfall condition to its associated k value. During model execution, the algorithm uses this reference dataset alongside the characteristics of the incoming input sample to dynamically select and apply the most suitable k in real time.

This novel approach enables the model to generalize more effectively across various rainfall conditions, ensuring that the chosen k-value is well-suited to the given rainfall regime. The results from this procedure showed that lower values of k capture noise and exhibit signs of overfitting the training data. However, higher values of k, especially where $k > 20$, tend to smooth predictions and result in underfitting of the dataset. As a result, these values were discarded.

4.4. Analysis of Optimal k-Values for Various Rainfall Events

The analysis of the optimal k-value in relation to rainfall rate, R, across different rainfall regimes from the observed data reveals distinct patterns that vary based on the nature of precipitation. These patterns are learned and subsequently influence the adaptive behaviour of the model in choosing the optimal k-value, enabling it to adapt dynamically to various rainfall intensities across different events. Fig. 1 illustrates the optimal k-value versus rain rate patterns identified for different rainfall regimes during the training process.

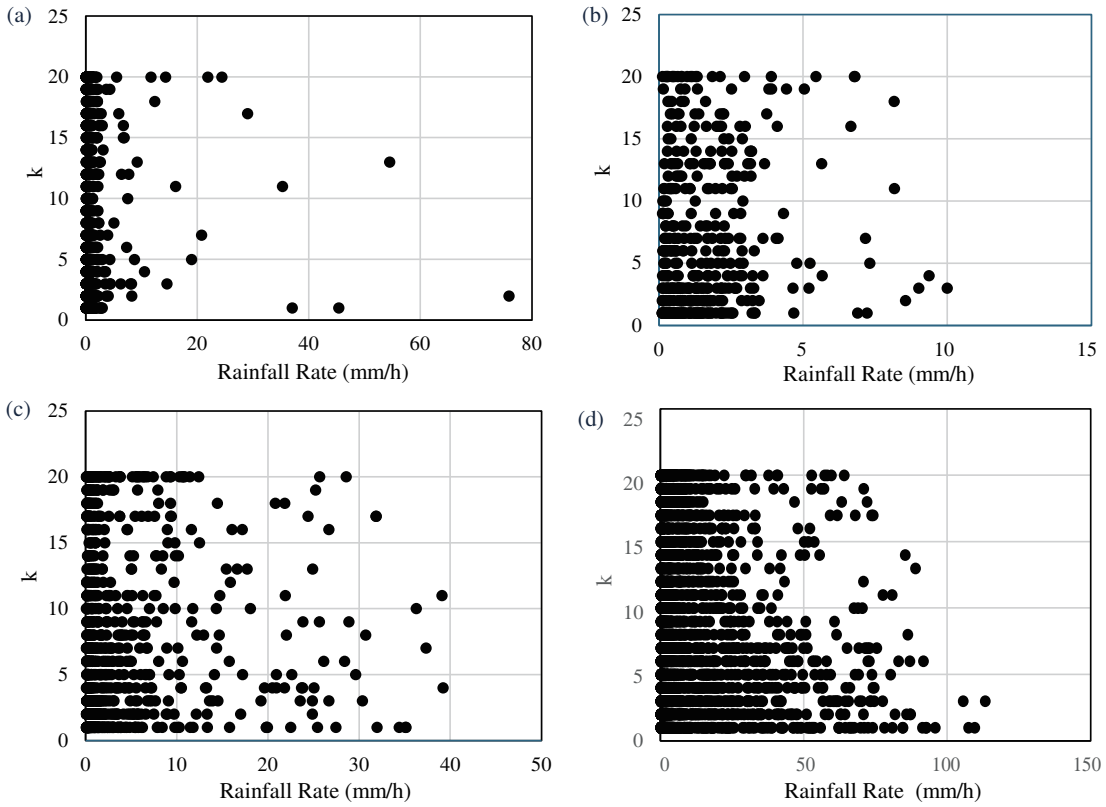

FIGURE 1. Optimal k analysis in (a) Drizzle, (b) Widespread, (c) Shower and, (d) Thunderstorm regimes.

In drizzle events, no strong linear correlation exists between k and R, as the optimal k-values appear scattered, indicating their dependence on local variations in DSD. However, a trend emerges where smaller k-values between 1 and 6 are more frequent at higher rainfall rates above 1 mm/h, suggesting that smaller neighbourhoods provide better local estimations due to increased DSD variability. Conversely, for lower rainfall rates less than 0.5 mm/h, larger k-values between 10 and 20 are present, likely aiding in smoothing out local fluctuations.

In widespread rainfall regimes, no clear monotonic trend exists, but for low values of R less than 1 mm/h, k-values range widely, indicating the high variability of DSD. Moderate rainfall rates between 1 and 5 mm/h exhibit more stable k-values, frequently between 10 and 17, while higher rainfall rates above 5 mm/h favour higher k-values between 16 and 20. Unlike drizzle, widespread rainfall does not show abrupt shifts in k-values but follows a gradual transition.

The scatter plot reveals a correlation between R and the optimal k-value in shower rainfall events. Although a strict linear correlation cannot be established, trends do appear across various rain rate ranges. At low rain rates, higher k-values ($k = 7$, 11, 12, 16, 17, 18, 20) are more common, and at rain rates higher than 10 mm/h, k is generally larger, often reaching 20, which indicates the need for broader neighbourhoods to sustain prediction stability. The horizontal axis follows a logarithmic scale, highlighting a nonlinear relationship where k increases with rain rate, but not in a proportional manner. The visualization typically suggests a moderate positive correlation, where higher rain rates tend to correspond to higher optimal k-values, albeit with some variability at lower rain rates.

For thunderstorm events, the relationship between optimal k-values and rainfall rate tends to be non-linear, suggesting complex interactions between rainfall intensity and DSD characteristics. Like shower rain events, thunderstorm events demonstrate a weak correlation between rainfall rate and k, with highly variable k-values at different rainfall levels. The scatter plot analysis suggests that extreme conditions (very low and very high rainfall) tend to have smaller k-values, possibly due to rapid changes in DSD. The presence of logarithmic trends in the rainfall rate further supports the need for dynamic tuning in KNN models.

Overall, the findings highlight the necessity of an adaptive KNN approach, as rainfall-induced attenuation modelling cannot rely on a fixed k-value, but should instead be adjusted based on the specific characteristics of each rainfall regime.

4.5. Random Forest (RF) Approach for DSD Estimation

The RF algorithm employs an ensemble learning approach for building various decision trees and merging their results to enhance accuracy, minimize errors and overfitting. This study uses the RF model to predict the DSD for each drop diameter bin, following the approach suggested as follows [31]:

$$N_D(X) = \frac{1}{M} \sum_{m=1}^{M} f_m(X) \tag{21}$$

where M is the total count of decision trees included in the ensemble, and $f_m(X)$ represents the prediction made by the m^{th} decision tree for the input X. To reduce prediction error in DSD estimation, the model adopts a performance criterion based on the MSE, as described by Hastie et al. [32],

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}\left[N(D)_{true} - N(D)_{pred}\right]^2 \tag{22}$$

This MSE metric evaluates the deviation of predicted DSD from actual values, aiding model optimization. The RF model was implemented using Jupyter Notebook, utilizing the *Random Forest Regressor* module from the scikit-learn for regression library.

4.6. Approach to Using Decision Trees for DSD Modelling

Decision tree (DT) is a rule-based model that recursively splits the dataset based on attribute values to create a predictive structure. For DSD prediction, the DT model estimates $N_D(X)$ by averaging the target values of the training samples that fall into the same terminal node (or region) R_j. This approach, originally proposed by Quinlan [24], is expressed as:

$$N_D(X) = \frac{1}{N_j}\sum_{i\epsilon R_j} N_D(X_i) \tag{23}$$

where N_j represents the count of training samples within region R_j, and $N_D(X_i)$ is the observed value corresponding to the i^{th} sample.

In this study, the DT model is initialized via the *DecisionTreeRegressor* class from the Scikit-learn library, setting the *random_state* parameter to a constant value for consistent results. Setting this parameter guarantees that the same sequence of random numbers is used each time the model is run, making the outcomes reproducible.

4.7. Training the DT Model

The model undergoes training on the dataset by gradually splitting the data, $D = \{X_1, X_2, X_3 \ldots, X_n\}$, at each node according to a chosen feature f_j and its corresponding threshold value θ_j. During this process, the algorithm selects the best feature for splitting at each node. Accordingly, the dataset is then divided into two subsets at each split [24]:

$$D_t^{left} = \{X_i | f_j(X_i) \le \theta_j\} \tag{24a}$$

and,

$$D_t^{right} = \{X_i | f_j(X_i) > \theta_j\} \tag{24b}$$

where $f_j(X_i)$ denotes the j^{th} feature corresponding to the data point X_i, and θ_j is the selected threshold for that feature. The best feature f_j and threshold θ_j are selected by minimizing a splitting criterion — specifically, the MSE in this study. The MSE for each possible split is calculated by adding the squared differences between the actual and predicted target values for both subsets [24, 32, 33],

$$\text{MSE}_t = \frac{1}{D_t^{left}}\sum_{i\in D_t^{left}}(y_i - \hat{y}_i)^2 + \frac{1}{D_t^{right}}\sum_{i\in D_t^{right}}(y_i - \hat{y}_i)^2 \tag{25}$$

Then, the optimal feature f_j and threshold θ_j combination is selected from the lowest calculated MSE using the equation:

$$(f_j^*, \theta_j^*) = \arg\min_{f_j,\theta_j} \text{MSE}_t \tag{26}$$

This equation leads to the most accurate regression model at each split in the decision tree [24, 32].

4.8. Model Performances with Metric Measures

This study employed three well-established regression evaluation metrics — Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Coefficient of Determination (R^2) — to objectively assess the performance of the machine learning models in predicting DSD. These metrics are standard in predictive modelling literature and offer complementary insights into model accuracy, residual behaviour, and overall predictive power [34, 35].

The RMSE serves as a widely recognised metric that measures the standard deviation of prediction errors, and it is defined as [34],

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{27}$$

Here, y_i represents the observed value, $\hat{y}_i$ the anticipated value, and n the total count of observations. A lower RMSE suggests a better model performance. The MAE calculates the mean magnitude of errors in a set of predictions, ignoring their direction, and is given by,

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{28}$$

It is less impacted by outliers than RMSE and offers a clear understanding of average prediction error [35].

The R^2 score, or coefficient of determination, shows how much variance in the dependent variable can be predicted from the independent variables. It is a standardised measure of model fit, defined as [36],

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \tag{29}$$

where $\bar{y}_i$ the mean of the observed values is located. A higher R^2 value, closer to 1, suggests a strong correlation between predicted and actual outcomes, indicating that the model explains a substantial amount of the variance. Negative values implies a poor model performance in comparison to the mean of the observed data.

In addition to offering single-value estimates of model accuracy, the interpretation of RMSE, MAE, and R^2 can be enhanced by using confidence intervals (CIs). Confidence intervals help quantify the degree of uncertainty associated with each metric, providing an estimated range in which the true population parameter is likely to fall at a specified confidence level, typically 95%. A confidence interval is generally expressed in the form [37]:

$$CI = \hat{\theta} \pm Z_{\alpha/2} \times SE\left(\hat{\theta}\right) \tag{30}$$

where $\hat{\theta}$ is the sample estimate (e.g., mean R^2, RMSE, or MAE) and $Z_{\alpha/2}$ is the critical value from the standard normal distribution corresponding to the desired confidence level. The standard error of the estimate, $SE(\hat{\theta})$, is calculated as:

$$\text{SE}\left(\hat{\theta}\right) = \frac{\sigma}{\sqrt{n}} \tag{31}$$

where σ is the sample standard deviation of the metric across the rainfall events, and n is the number of rainfall cases included in the evaluation of each metric.

Narrower intervals indicate greater precision and reliability, while wider intervals suggest higher variability and less certainty in the estimate [37, 38]. In this study, the reporting of 95% CIs for RMSE, MAE, and R^2 not only enables a more rigorous comparison of model performance but also ensures that conclusions drawn are statistically robust and reliable.

4.9. Estimation of Rainfall Attenuation

Rain-induced attenuation of electromagnetic waves is estimated using DSD data alongside Mie scattering theory, which provides a comprehensive representation of wave interactions with raindrops from microwave to terahertz frequencies. Rainfall causes signal attenuation primarily due to the scattering and absorption of electromagnetic waves by raindrops [39]. The methodology involves three computational steps: determining the complex index of refraction of water, calculating the extinction cross-section (ECS), and estimating specific attenuation. The exact reduction in radio wave strength caused by rainfall is determined using a modified version of the formula in Adimula and Ajayi [3] as further proposed in [9, 29]:

$$A_s = 4.343 \times 10^{-3} \sum_{n=1}^{20} N(D_n) k_{ext} a^{\varsigma_{ext}} dD_n \text{ [dB/km]} \tag{32}$$

The specific attenuation, A_s, is directly affected by the distribution of raindrop sizes $N(D_n)$ and the coefficients of ECS; k_{ext} and ς_{ext}. The ECS measures the overall impacts of scattering and absorption of electromagnetic waves by raindrops. The values of ECS are analyzed for different drop sizes at various frequencies and ambient temperatures to evaluate how scattering and absorption affect electromagnetic signal propagation in rainfall conditions. To determine the scattering parameters, Mie coefficients and Bessel functions are employed. The ECS, denoted as Q_{ext}, is calculated using the formulation provided by Hult and Van de Hulst [40]:

$$Q_{ext}\left(D\right) = \frac{4\pi}{k^2} \text{Re}\left\{s\left(0\right)\right\} \tag{33}$$

where k is the wavenumber corresponding to the frequency of the wave in a rain-filled medium, and $s\left(0\right)$ represents the real component of the forward scattering amplitude, as defined by Mie [41]:

$$s\left(0\right) = \frac{1}{2} \sum_{n=0}^{\infty} \left(2n+1\right) \left[a_n\left(m,\alpha\right) + b_n\left(m,\alpha\right)\right] \tag{34}$$

In this summation, a_n and b_n correspond to the coefficients derived from Mie scattering theory, where m represents the complex refractive index of water. The frequency of raindrops and the surrounding temperature significantly impact how the α parameter changes. The average annual temperature of 20.27°C was estimated for the location of Durban between 2018 and 2019. The computation results are undertaken to reveal frequency-dependent trends in specific attenuation modelling. Therefore, ECS was calculated for frequency range up to 100 GHz, covering both microwave and millimetre-wave bands.

To evaluate the performance of ML-based attenuation models, predictions are compared against both traditional statistical models and ITU-R recommendations under various rainfall regime variations.

5. RESULTS AND DISCUSSION

5.1. Performance Evaluation of Machine Learning Models

This section assesses the performance of k-nearest neighbours (KNN), decision trees (DT), and random forest (RF) models in predicting DSDs across various rainfall regimes: drizzle, widespread rain, showers, and thunderstorms. For each model, its capability to capture DSD variations at different rainfall intensities is evaluated, highlighting their adaptability and generalization capabilities, as presented in Figure 2.

During drizzle events, KNN closely approximates $N(D)$ for mid-range drop diameters but tends to overestimate at lower diameters. The DT algorithm performs well in the mid-diameter range but tends to underestimate small drops and overestimate large ones. The RF approach performs well for small to mid-sized drops but deviates for larger diameters. Figure 2(a) shows the evaluation of the models during a drizzle rain event at 1.33 mm/h, where raindrop variability is minimal. KNN performed the best by executing the local averaging of the retained dataset to capture subtle changes in DSD. The decision tree, while capable of recognising general patterns, exhibited signs of overfitting due to its sensitivity to minor fluctuations in low-variability data. Random Forest, by averaging multiple decision trees, offers a smoother and more stable output, though it may miss some fine-grained DSD detail.

For a widespread rain event at 6.79 mm/h in Figure 2(b), KNN continued to perform well, particularly with adaptive tun-

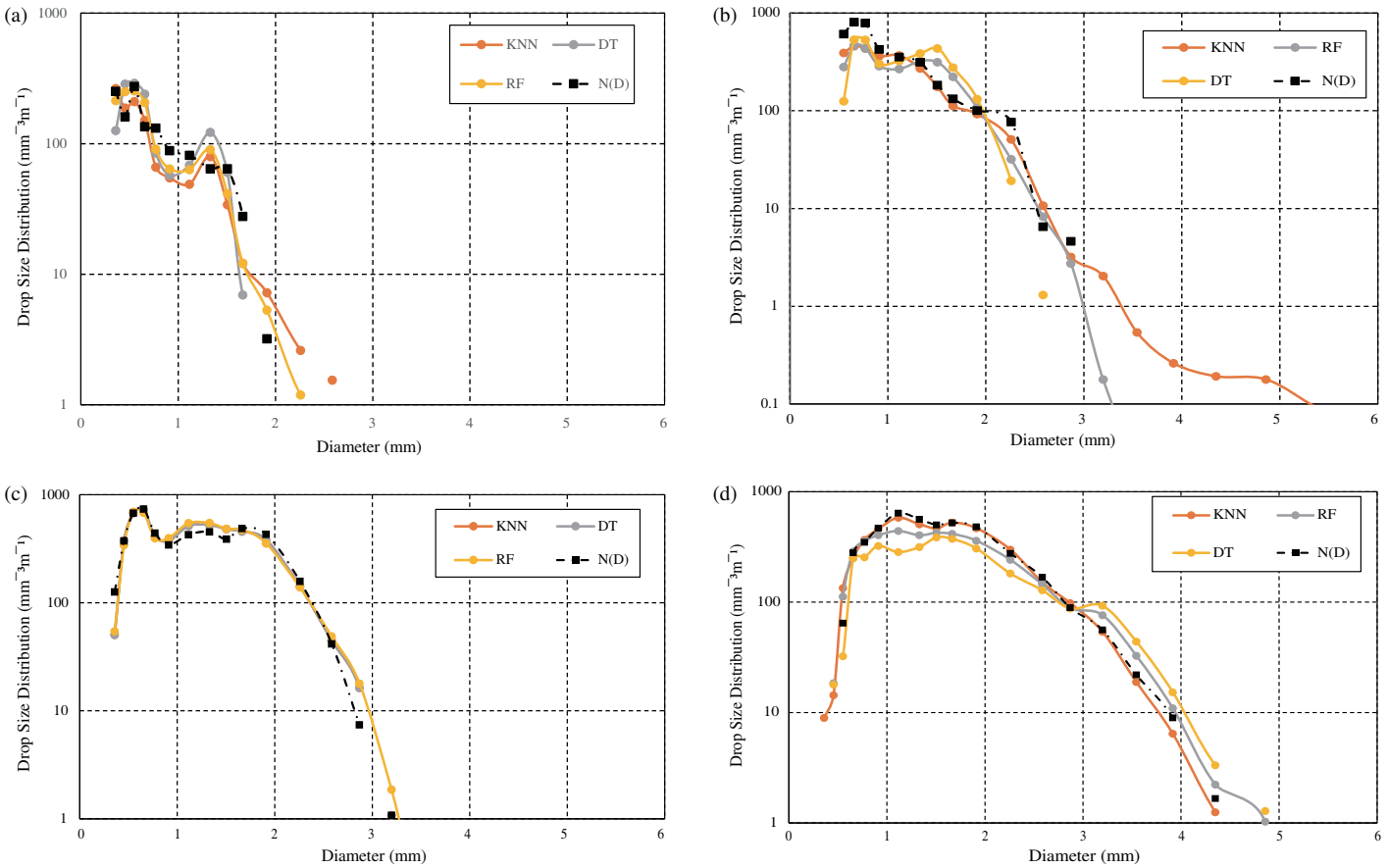

FIGURE 2. Evaluation of machine learning models for DSD prediction across various rainfall regimes: (a) Drizzle at 1.33 mm/h, (b) Widespread at 6.79 mm/h, (c) Shower at 34.39 mm/h, (d) Thunderstorm at 81.61 mm/h.

ing of the neighbourhood size with DT showing reduced generalization. Meanwhile, RF produced robust and accurate predictions due to its ensemble nature, effectively handling the moderate complexity in the data.

In the shower regime, the selected rainfall rates for consideration range from 14.67 mm/h to 34.39 mm/h. KNN provides the most accurate estimates across the range, closely tracking observed $N(D)$ values, particularly in mid-sized drops as seen in Figure 2(c). RF exhibits variability, performing well while estimating smaller diameters, but often overestimating mid-sized drops. DT tends to underperform, particularly for higher rainfall rates, with frequent underestimations in mid-to-large drop sizes. At 34.39 mm/h, all three models showed nearly identical performance across the range of drop sizes while underestimating smaller drops and overestimating drops larger than 2.5 mm.

During thunderstorms at a rain rate of 81.61 mm/h as shown in Figure 2(d), both RF and DT showed significant performance degradation. The ensemble structure of RF was inadequate for the highly dynamic input, and DTs failed due to overfitting and limited depth. KNN remained the most reliable model in this regime, offering stable and accurate predictions thanks to adaptive tuning and local averaging, which demonstrated the ability to handle noise and complexity.

Overall, KNN proved to be the most robust and adaptable model across all rainfall regimes, particularly excelling in high-intensity cases conditions. Random forest performed well in low to moderate rainfall scenarios, while the DT model was generally less suitable for complex or noisy datasets.

5.2. Error Analysis of DSD Modelling for Different Rainfall Regimes

A comparative analysis of DSD modelling across different rainfall regimes — drizzle, widespread rain, showers, and thunderstorms — using KNN, RF, and DT reveals that KNN consistently outperforms the other models. In Table 1, KNN achieves the highest R^2 score values (with an average of 0.92), indicating strong correlation with observed $N(D)$ values, while RF ($R^2 = 0.85$) and DT ($R^2 = 0.77$) show more variability. In terms of these error metrics, KNN maintains the lowest RMSE score of 43.61 in Table 2 and MAE score of 26.58 in Table 3, ensuring the most precise predictions across different drop sizes and rainfall intensities. RF, while showing some promise, exhibits higher prediction errors, especially in mid-sized drop distributions, while DT consistently underperforms, particularly in higher rainfall rates where its RMSE and MAE values are significantly larger. Across all rainfall regimes, KNN provides the most reliable estimates, demonstrating lower RMSE and MAE, even at extreme high rainfall rates such as 94.4 mm/h. RF occasionally competes with KNN but shows greater fluctuations,

TABLE 1. Coefficient of determination (R^2 score) for DSD predictions across different rainfall regimes using ML models.

Rainfall Category	R (mm/h)	KNN	RF	DT
Drizzle	0.814	0.92	0.85	0.55
	1.33	0.9	0.86	0.64
	1.16	0.89	0.73	0.67
	1.07	0.9	0.82	0.64
Widespread	1.63	0.89	0.87	0.78
	1.39	0.91	0.89	0.91
	9.99	0.85	0.88	0.72
	8.54	0.88	0.86	0.77
Showers	14.66	0.94	0.67	0.47
	34.39	0.97	0.95	0.97
	19.83	0.94	0.74	0.94
	24.87	0.94	0.93	0.91
Thunderstorm	60.33	0.94	0.92	0.89
	81.61	0.98	0.89	0.70
	62.77	0.98	0.96	0.94
	94.4	0.89	0.84	0.89
Average		0.92	0.85	0.77

TABLE 2. Root mean square error (RMSE) for DSD predictions across different rainfall regimes using KNN, RF, and DT models.

Rainfall Category	R (mm/h)	KNN	RF	DT
Drizzle	0.814	13.86	19.2	33.82
	1.33	25.77	30.88	49.94
	1.16	13.04	20.45	22.87
	1.074	16.09	22.04	31.67
Widespread	1.63	41.12	44.77	58.52
	1.39	41.95	47.23	41.95
	0.42	42.09	38.50	58.37
	8.54	65.35	70.85	92.48
Showers	14.66	66.3	165.82	210.17
	34.39	40.39	51.44	40.40
	19.83	45.05	94.16	45.05
	24.87	76.16	79.36	94.05
Thunderstorm	60.33	49.07	56.95	65.21
	81.61	27.69	73.00	122.38
	62.77	31.9	42.67	54.51
	94.4	102	124.08	102.84
Average		43.61	61.34	70.26

TABLE 3. Mean absolute error (MAE) for DSD predictions across different rainfall categories using KNN.

Rainfall Category	R (mm/h)	KNN	RF	DT
Drizzle	0.814	6.78	8.01	19.44
	1.33	16.07	18.61	27.28
	1.16	8.02	11.88	12.59
	1.074	8.34	11.97	15.8
Widespread	1.63	19.85	27.02	34.72
	1.39	20.42	26.75	20.42
	0.42	25.5	21.84	35.58
	8.54	43.28	44.13	61.18
Showers	14.66	40.12	94.1	25.22
	34.39	25.22	35.55	29.35
	19.83	29.35	51.02	51.1
	24.87	44.19	44.54	89.59
Thunderstorm	60.33	32.32	35.88	40.54
	81.61	18.59	45.32	78.53
	62.77	20.09	25.77	34.46
	94.4	67.25	74.31	67.25
Average		26.58	36.04	40.19

whereas DT model suffers from higher errors and lower predictive accuracy.

To provide a statistically grounded comparison, 95% confidence intervals (CIs) were estimated for each metric across the rainfall regimes, as summarized in Table 4. The coefficient of determination (R^2) indicates the strength of the correlation between the KNN model and the measured DSD. The KNN model

TABLE 4. Performance metrics with 95% confidence intervals (CI) for DSD predictions using KNN, RF, and DT models.

Metric	Model	95% CI
R^2	KNN	0.92 ± 0.02
	RF	0.85 ± 0.04
	DT	0.77 ± 0.08
RMSE	KNN	43.61 ± 12.90
	RF	61.34 ± 21.25
	DT	70.26 ± 24.91
MAE	KNN	26.59 ± 8.52
	RF	36.04 ± 12.31
	DT	40.19 ± 12.28

achieves an impressive average R^2 value of 0.92 ± 0.02, validating over 90% of the variance in the observed data with minimal uncertainty. In comparison, the RF model reaches R^2 of 0.85 ± 0.04, while the DT model shows the weakest fit with an R^2 of 0.77 ± 0.08. The relatively narrow confidence interval for the KNN model underscores its consistent predictive capability across various rainfall conditions. In contrast, the broader intervals for the RF and DT models suggest greater variability and less stable performance.

When evaluating error margin, KNN again demonstrates superior performance, achieving the lowest RMSE margin of 43.61 ± 12.90. In contrast, RF produces a higher margin of 61.34 ± 21.25, while DT records the largest margin at 70.26 ± 24.91. The comparatively wider confidence intervals for RF and DT reflect greater variability and reduced reliability in predicting the DSD across different rainfall conditions, particu-

TABLE 5. Input parameters obtained for DSD models using MoM technique in Durban.

DSD Model	Input Parameters	Coefficients		R^2
Lognormal	N_t	$a = 102.42$	$b = 0.51$	0.4199
	μ	$A_\mu = -0.073$	$B_\mu = 0.127$	0.4296
	σ^2	$A_\sigma = 0.0795$	$B_\sigma = 0.0023$	0.0097
Gamma ($\mu = 2$)	N_0	$a = 13327$	$b = 0.0911$	0.0087
	$\wedge$	$a_\Lambda = 4.8839$	$b_\Lambda = -0.135$	0.4144

larly during convective events. By comparison, the narrower CI associated with KNN underscores both its accuracy and stability in error performance.

The MAE results further highlight the superiority of KNN, which achieved an average of 26.59 ± 8.52. This significantly outperformed both RF, with an average of 36.04 ± 12.31, and DT, which had an average of 40.19 ± 12.28. These findings indicate that KNN not only produces smaller deviations from the measured values but also maintains a consistent level of accuracy across various rainfall conditions.

In summary, the error analysis establishes KNN as the most robust and dependable regression model for DSD prediction. Its statistical reliability, verified through 95% confidence intervals, underscores its suitability for rainfall attenuation modelling and related hydrometeorological applications.

5.3. Comparison of ML DSD Model and Traditional Statistical DSD Models

The KNN model, identified as the top-performing machine learning model, is assessed and contrasted against the performance of two selected traditional statistical methods. The comparison is based on observed variations of $N(D)$ values with different rainfall rates, R, specifically focusing on the strengths of each model when predicting DSDs across varying drop diameters. These models were selected because several authors have demonstrated over the years that they provide a good representation of DSD variations in Durban and other subtropical regions on the continent [9], [19, 20]. The modified gamma and lognormal distribution size models were chosen to represent the drop size distribution across various rainfall rates in Durban.

The method of moments (MoM) technique is applied to estimate DSD parameters for these models [6, 11]. The MoM derives these parameters from actual DSD measurements by calculating specific statistical moments and utilizing non-linear regression techniques to fit input parameter functions. Regression models obtained from MoM approximations for these two models are presented in Table 5 based on the input parameter functions given in (3) and (4).

Figure 3 illustrates a comparison of three models — k-Nearest Neighbours (KNN), lognormal (LogN), and modified gamma (Gamm) — for estimating the DSD under different rainfall conditions. Our outcomes show that KNN provides the most accurate depiction of the observed DSD, remaining closely aligned with the measured values for all tested rain rates.

In Figure 3(a), a light rain intensity of 5.07 mm/h shows that all three models approximate the empirical DSD reasonably well. However, KNN aligns most closely with the measured curve, especially around the peak drop concentration and modal drop size. The lognormal distribution tends to undercount smaller droplets (diameters less than 1 mm), a common limitation due to its symmetric form, which restricts its ability to capture the skewed behavior of natural drop size distributions, particularly in lighter rainfall conditions. The gamma model overpredicts across the entire diameter range, producing broader, exaggerated curves that deviate from the observed distribution. As rain intensity increases, as shown in Figures 3(b)–(e), these trends remain evident. During moderate rainfall intensities of 46.53 mm/h and 59.94 mm/h, the lognormal model tends to underestimate the concentrations of mid-sized and smaller droplets. In contrast, the gamma model often overestimates concentrations, particularly for larger diameters. Meanwhile, the K-Nearest Neighbours (KNN) model closely aligns with the empirical DSD curves, effectively capturing both the central peak and the tail behaviour with high accuracy

During heavy rainfall with rates exceeding 80 mm/h, the drop size distributions often display multimodal and highly skewed characteristics, which further highlight the limitations of statistical models. The KNN model, with its adaptive nearest-neighbour search, aligns well with the empirical curves and maintains fair accuracy throughout the entire range of drop diameters.

In summary, the lognormal model continues to show a downward bias, particularly for mid-sized and small drops. Meanwhile, the overestimation by gamma model becomes more pronounced at higher intensities, suggesting a mismatch in its parameterization under such conditions. Overall, the results underscore the limitations of traditional statistical models, which rely on fixed functional forms, making them less responsive to the variability inherent in real-world precipitation data.

KNN is a data-driven, non-parametric method that offers a more flexible and precise tool for modelling DSDs, especially over a broad spectrum of rainfall intensities.

5.4. Analysis of Specific Attenuation Across Rainfall Regimes

Variations in specific attenuation across different rainfall regimes offer valuable insights into the effectiveness of machine learning models compared to traditional statistical models, such as lognormal (LogN) and ITU-R P.838-3 pre-

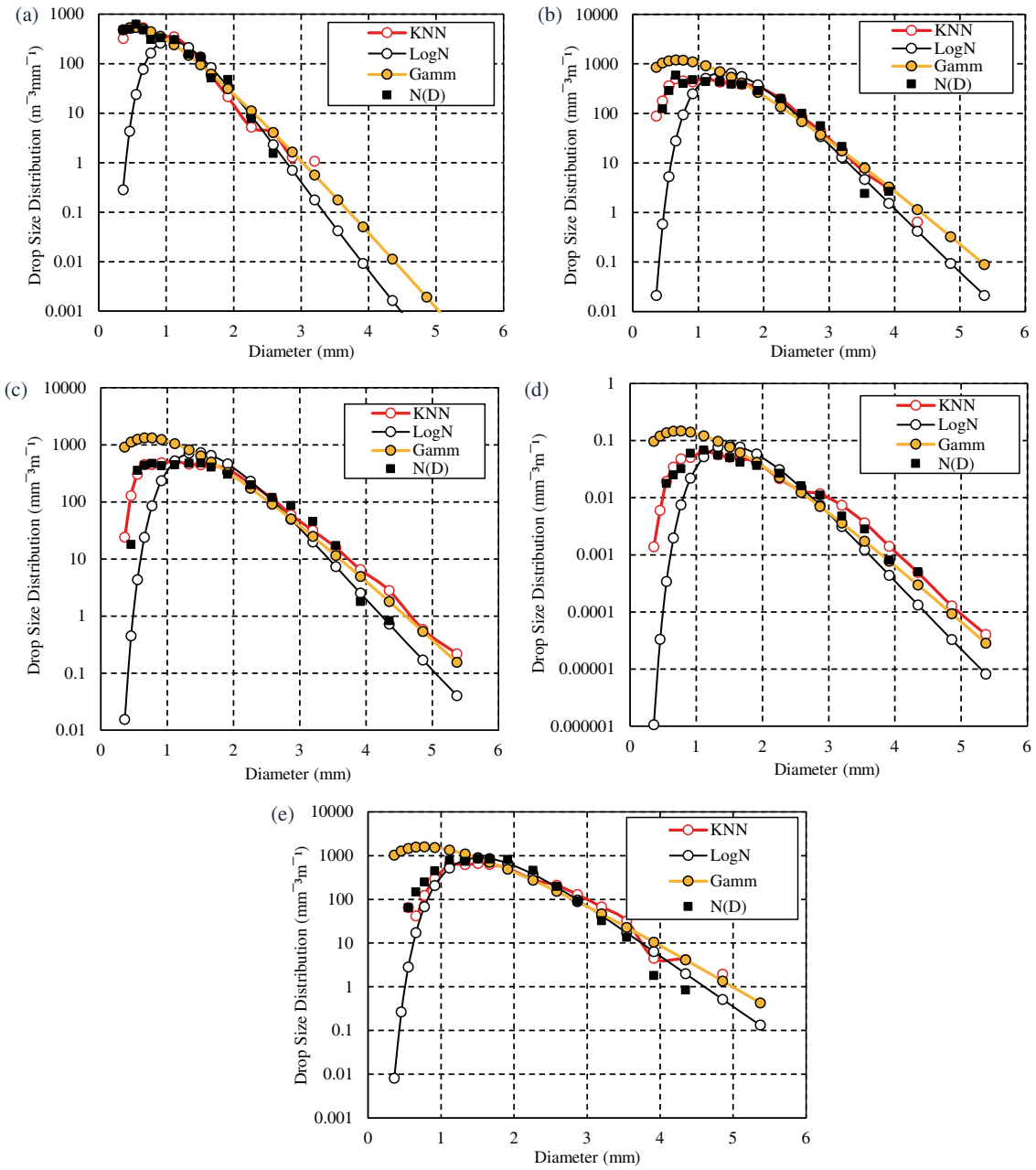

FIGURE 3. Comparison of ML models versus statistical models for DSD prediction at: (a) 5.07 mm/h, (b) 46.53 mm/h, (c) 59.94 mm/h, (d) 80.05 mm/h, (e) 98.02 mm/h.

diction models (horizontal and vertical) specified in [42]. For this analysis, a frequency range of 2 GHz to 100 GHz has been chosen, encompassing four distinct rainfall regimes: drizzle, widespread rain, showers, and thunderstorms, each characterized by different rainfall intensity rates.

For drizzle, with an average rainfall rate of $R = 1.23$ mm/h, the measured specific attenuation at lower frequencies less than 10 GHz is relatively low, around 0.01 dB/km as depicted in Figure 4. The machine learning models (KNN, RF, and DT) align closely with the observed values, with KNN showing a slight deviation of 0.0002 dB/km from the measured values, while RF and DT yields almost identical results. In contrast, the statistical models show minor deviations, with the lognormal and ITU-R (vertical) models slightly underestimating the specific attenuation at this low rainfall intensity. As the frequency increases, the discrepancy between the machine learning models and the statistical models becomes more evident. For example, at 50 GHz, the KNN model yields a value of 0.33 dB/km, closely aligning with the measured data, while the ITU-R models show increased discrepancies, indicating their reduced accuracy at higher frequencies during drizzle events.

During a widespread rain event with an intensity of 8.54 mm/h, as shown in Figure 5, the specific attenuation increases notably compared to the drizzle regime. For instance, at 2 GHz, the measured attenuation is 0.0015 dB/km. KNN delivered reliable results, predicting a value of 0.0013 dB/km,

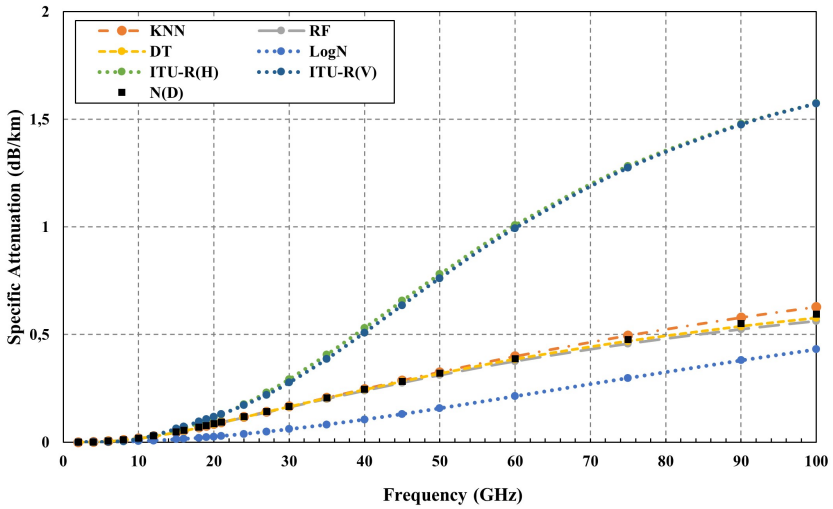

FIGURE 4. Compared specific attenuations estimated for drizzle events at rain rate of 1.23 mm/h.

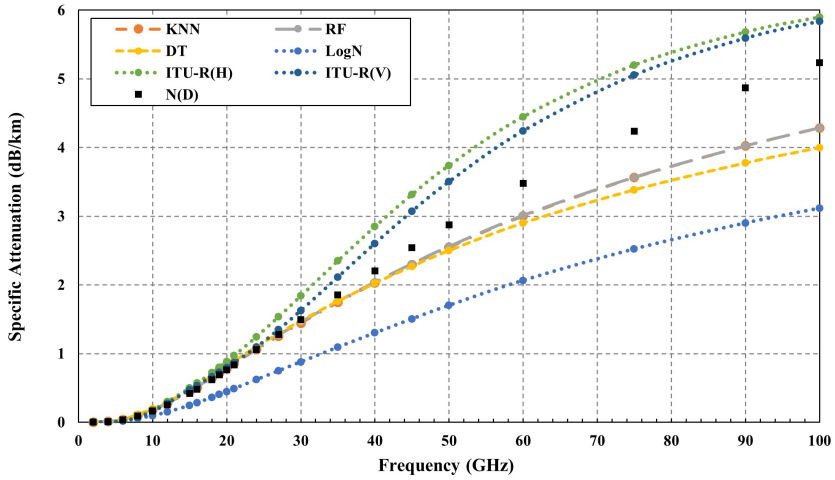

FIGURE 5. Compared specific attenuation estimated for widespread rainfall at a rain rate of 8.54 mm/h.

which is very close to the observed specific attenuation. RF and DT also perform favorably, with RF showing slight underestimation. In comparison, the statistical models display a trend of underestimating attenuation values, particularly the ITU-R models, which diverge more noticeably with increasing frequency. At frequencies of 60 GHz and 100 GHz, the models demonstrate a widening gap in their predictions, especially between machine learning and statistical models. The RF model predicts 9.39 dB/km at 100 GHz, whereas the ITU-R (vertical) model predicts 11.28 dB/km, indicating a noticeable overestimation by the ITU-R model.

For shower regime, where the rainfall intensity is 22.62 mm/h as presented in Figure 6, the recorded attenuation reaches 0.019 dB/km at 2 GHz. The machine learning models again demonstrate robust performance, with KNN and RF predicting values closely matching the measured attenuation. For instance, at 50 GHz, the KNN model estimates an attenuation of 5.58 dB/km, while the RF model gives 5.24 dB/km.

On the other hand, the statistical models — lognormal and ITU-R — consistently overestimate the attenuation within this frequency range. At 75 GHz, the specific attenuation for showers reaches 7.72 dB/km. The ML models tend to provide better estimates, with KNN and RF aligning more closely with the measured attenuation compared to the statistical models, which show a tendency to overpredict.

Thunderstorm events, with an average rainfall rate of 72.80 mm/h, produce the highest levels of rainfall, leading to a significant increase in specific attenuation as seen in Figure 7. At 2 GHz, the measured value of specific attenuation is 0.0084 dB/km, which increases rapidly with frequency. At higher frequencies, between 75 GHz and 100 GHz, the ML models predict attenuation values that closely match the observed data. For instance, at 100 GHz, KNN predicts a value of 21.75 dB/km, while the RF and DT models have respectively estimated 22.08 dB/km and 21.07 dB/km, showing close agreement with the actual measured value. ITU-R models, particularly ITU-R (Horizontal), show a significant

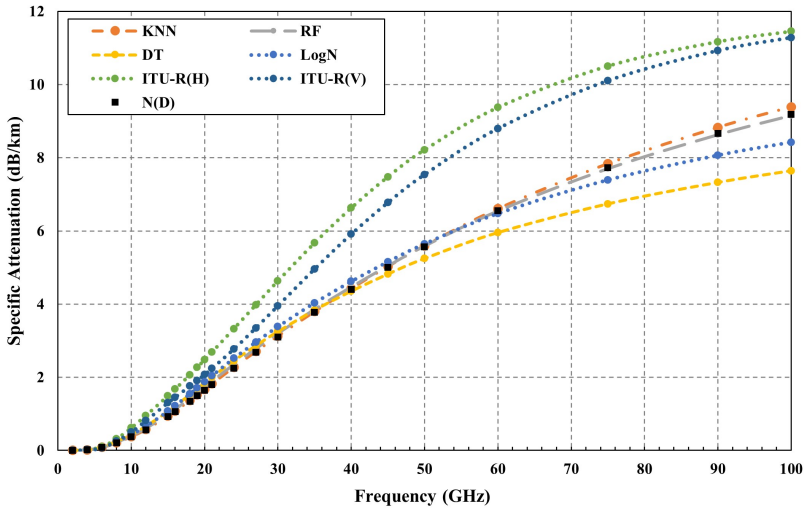

FIGURE 6. Compared specific attenuation estimated for shower events at a rain rate of 22.62 mm/h.

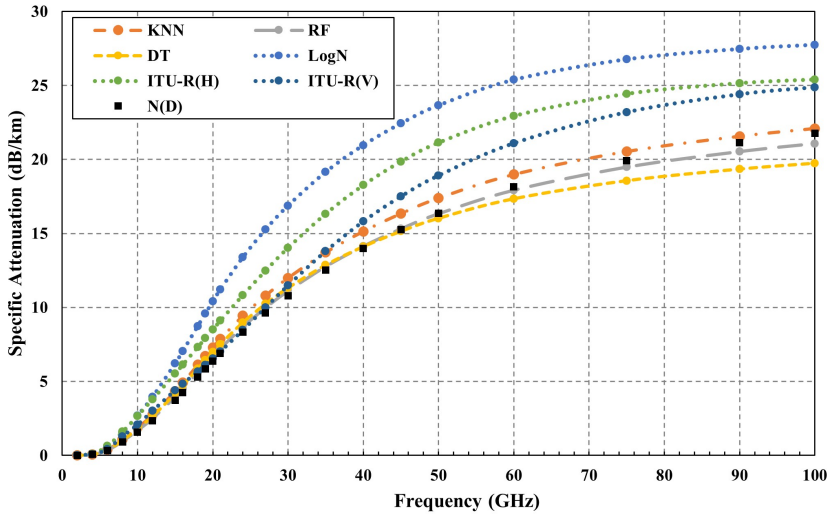

FIGURE 7. Compared specific attenuation estimated for thunderstorm events at 72.80 mm/h.

overestimation of attenuation at frequencies beyond 50 GHz. For instance, at 100 GHz, the ITU-R (horizontal) model predicts 27.74 dB/km, much higher than the expected measured value. This overestimation is especially pronounced at high frequencies, demonstrating the limitations of the traditional models under intense rainfall conditions.

Results from our study demonstrate that machine learning models (KNN, RF, and DT) outperform traditional statistical models (lognormal and ITU-R) across the four rainfall regimes. These models excel at capturing the increase in specific attenuation at higher frequencies, particularly in the widespread rain and thunderstorm regimes. KNN stands out with the best overall performance, providing accurate predictions that closely match observed data across various rainfall intensities and frequencies. While RF and DT also produce good results, the ability of KNN to adapt to different rainfall conditions, especially at higher frequencies, highlights its superiority. In contrast, statistical models show limited effectiveness, performing better in the drizzle and widespread rain regimes but exhibiting greater discrepancies as rainfall intensity and frequency rise. The ITU-R models tend to overestimate specific attenuation in Durban, particularly within the thunderstorm regime at higher frequencies, revealing their decreasing accuracy under extreme rainfall conditions. Overall, the results emphasize that machine learning models, especially KNN, are more adaptable to varying rainfall intensities and frequencies, delivering more accurate and reliable predictions of specific attenuation. This suggests that KNN is a suitable ML tool for regions with variable rainfall patterns, such as subtropical climates, enabling more efficient radio link design and maintenance.

Comparing the results of this study with previous studies in [2], [17], and [19] reveals a persistent issue: ITU-R models overestimate specific attenuation due to rainfall at the measurement site. Although path attenuation is not computed, as rain cell estimation needs to be considered, it is evident that machine learning models in subtropical regions offer more accurate esti-

mations than the widely used lognormal DSD model. As wireless systems approach the global implementation of 6G/7G, where millimeter-wave bands are increasingly important, accurate link budget estimation for outdoor radio planning in turbulent wireless media, such as rainfall, could benefit from ML DSD models. As results have shown across various rainfall regimes, rainfall attenuation estimation from ML models will lead to improved power management for frequency-selective fade mitigation across the radio spectrum, particularly where ITU-R inadvertently overestimates the link margin.

6. CONCLUSION

This study has demonstrated that machine learning models, particularly KNN, offer a more accurate and adaptable approach to predicting rainfall attenuation than traditional empirical models, with RF as a strong alternative. While ITU-R and lognormal models provide baseline estimations, their reliance on simplified assumptions limits their applicability, especially in regions with complex and highly variable rainfall patterns such as Durban. The adoption of data-driven models in radio propagation planning can enhance network resilience, optimize performance, and ensure reliable communication under diverse rainfall conditions. With the continuous expansion of high-frequency communication systems (5G, 6G, and beyond), the integration of ML-based attenuation models will be essential in developing robust, efficient, and weather-resilient wireless networks. This study has significantly improved the accuracy of DSD modelling and attenuation predictions, contributing to better link budgeting and enhanced network reliability.

While the study introduced an adaptive k-selection method to enhance the performance of KNN model, the random forest (RF) and decision tree (DT) models were used with fixed hyperparameters. Future research will focus on implementing an adaptive parameter tuning method for these models, enabling them to automatically adjust hyperparameters based on dataset characteristics to improve their predictive accuracy, reduce overfitting, and increase their flexibility across different rainfall regimes.

ACKNOWLEDGEMENT

The authors hereby acknowledge the valuable contributions of Professor Thomas Afullo of the University of KwaZulu-Natal, Durban, South Africa with respect to the provision of Joss-Waldvogel disdrometer datasets used in this study.

REFERENCES

[1] Odedina, M. O. and T. J. Afullo, "Determination of rain attenuation from electromagnetic scattering by spherical raindrops: Theory and experiment," *Radio Science*, Vol. 45, No. 01, 1–15, 2010.

[2] Alonge, A. A. and T. J. O. Afullo, "Seasonal analysis and prediction of rainfall effects in Eastern South Africa at microwave frequencies," *Progress In Electromagnetics Research B*, Vol. 40, 279–303, 2012.

[3] Adimula, I. A. and G. O. Ajayi, "Variations in raindrop size distribution and specific attenuation due to rain in Nigeria," *Annals of Telecommunications*, Vol. 51, No. 1, 87–93, 1996.

[4] Timothy, K. I., J. T. Ong, and E. B. L. Choo, "Raindrop size distribution using method of moments for terrestrial and satellite communication applications in Singapore," *IEEE Transactions on Antennas and Propagation*, Vol. 50, No. 10, 1420–1424, Oct. 2002.

[5] Ulbrich, C. W., "Natural variations in the analytical form of the raindrop size distribution," *Journal of Climate and Applied Meteorology*, Vol. 22, No. 10, 1764–1775, 1983.

[6] Ajayi, G. O. and R. L. Olsen, "Modeling of a tropical raindrop size distribution for microwave and millimeter wave applications," *Radio Science*, Vol. 20, No. 02, 193–202, 1985.

[7] Owolawi, P. A., "Raindrop size distribution model for the prediction of rain attenuation in Durban," in *Progress In Electromagnetics Research Symposium Proceedings*, 1068–1074, Suzhou, China, 2011.

[8] Alonge, A. A. and T. J. Afullo, "Rainfall microstructures for microwave and millimeter wave link budget at tropical and subtropical sites," in *2013 Africon*, 1–5, Pointe aux Piments, Mauritius, 2013.

[9] Alonge, A. A., "Semi-empirical characteristics of modified lognormal DSD inputs using rain rate distributions for radio links over the African continent," *Advances in Space Research*, Vol. 67, No. 1, 179–197, 2021.

[10] Sarker, I. H., "Machine learning: Algorithms, real-world applications and research directions," *SN Computer Science*, Vol. 2, No. 3, 160, 2021.

[11] Kozu, T. and K. Nakamura, "Rainfall parameter estimation from dual-radar measurements combining reflectivity profile and path-integrated attenuation," *Journal of Atmospheric and Oceanic Technology*, Vol. 8, No. 2, 259–270, 1991.

[12] Tokay, A. and D. A. Short, "Evidence from tropical raindrop spectra of the origin of rain from stratiform versus convective clouds," *Journal of Applied Meteorology and Climatology*, Vol. 35, No. 3, 355–371, 1996.

[13] Testud, J., S. Oury, R. A. Black, P. Amayenc, and X. Dou, "The concept of "normalized" distribution to describe raindrop spectra: A tool for cloud physics and cloud remote sensing," *Journal of Applied Meteorology and Climatology*, Vol. 40, No. 6, 1118–1140, 2001.

[14] Altman, N. S., "An introduction to kernel and nearest-neighbor nonparametric regression," *The American Statistician*, Vol. 46, No. 3, 175–185, 1992.

[15] Breiman, L., J. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*, Chapman and Hall/CRC, 2017.

[16] Ramatladi, T. and A. Alonge, "Machine learning techniques for evaluating disdrometer-derived raindrop measurements over radio links," in *2023 IEEE AFRICON*, 1–6, Nairobi, Kenya, 2023.

[17] Afullo, T. J. O., "Raindrop size distribution modeling for radio link design along the eastern coast of South Africa," *Progress In Electromagnetics Research B*, Vol. 34, 345–366, 2011.

[18] Marshall, J. S. and W. M. K. Palmer, "The distribution of raindrops with size," *Journal of Meteorology*, Vol. 5, No. 4, 165–166, 1948.

[19] Malinga, S. J. and P. A. Owolawi, "Obtaining raindrop size model using method of moment and its applications for South Africa radio systems," *Progress In Electromagnetics Research B*, Vol. 46, 119–138, 2013.

[20] Alonge, A. A., "Correlation of rain dropsize distribution with rain rate derived from disdrometers and rain gauge networks in Southern Africa," MSc dissertation, Masters Dissertation submitted to the University of KwaZulu-Natal, Durban, 2011.

[21] Cover, T. and P. Hart, "Nearest neighbor pattern classification," *IEEE Transactions on Information Theory*, Vol. 13, No. 1, 21–27, 1967.

[22] Cho, P., M. Lee, and W. Chang, "Instance-based entropy fuzzy support vector machine for imbalanced data," *Pattern Analysis and Applications*, Vol. 23, No. 3, 1183–1202, 2020.

[23] Beyer, K., J. Goldstein, R. Ramakrishnan, and U. Shaft, "When is "nearest neighbor" meaningful?" in *International Conference on Database Theory*, 217–235, 1999.

[24] Quinlan, J. R., "Induction of decision trees," *Machine Learning*, Vol. 1, No. 1, 81–106, 1986.

[25] Breiman, L., "Random forests," *Machine Learning*, Vol. 45, No. 1, 5–32, 2001.

[26] Mallala, B., A. I. U. Ahmed, S. V. Pamidi, M. O. Faruque, *et al.*, "Forecasting global sustainable energy from renewable sources using random forest algorithm," *Results in Engineering*, Vol. 25, 103789, 2025.

[27] Louppe, G., "Understanding random forests: From theory to practice," Universite de Liege (Belgium) , 2004.

[28] Bartholomew, M. J., *Laser Disdrometer (LDIS) Instrument Handbook*, Oak Ridge National Laboratory (ORNL), Oak Ridge, TN (United States). Atmospheric Radiation Measurement (ARM) Data Center; Brookhaven National Laboratory (BNL), Upton, NY (United States), Tech. Rep., Jun. 2020. [Online]. Available: https://www.osti.gov/biblio/1226796

[29] Alonge, A. A. and T. J. Afullo, "Regime analysis of rainfall drop-size distribution models for microwave terrestrial networks," *IET Microwaves, Antennas & Propagation*, Vol. 6, No. 4, 393–403, 2012.

[30] Tokay, A., A. Kruger, W. F. Krajewski, P. A. Kucera, and A. J. P. Filho, "Measurements of drop size distribution in the southwestern Amazon basin," *Journal of Geophysical Research: Atmospheres*, Vol. 107, No. D20, LBA 19–1–LBA 19–15, 2002.

[31] Liaw, A. and M. Wiener, "Classification and regression by randomForest," *R News*, Vol. 2, No. 3, 18–22, 2002.

[32] Hastie, T., R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, ser. Springer series in statistics, Springer, 2009. [Online]. Available: https://books.google.com.hk/books?id=eBSgoAEACAAJ

[33] Loh, W.-Y., "Classification and regression trees," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, Vol. 1, No. 1, 14–23, 2011.

[34] Chai, T. and R. R. Draxler, "Root mean square error (RMSE) or mean absolute error (MAE)? — Arguments against avoiding RMSE in the literature," *Geoscientific Model Development*, Vol. 7, No. 3, 1247–1250, 2014.

[35] Willmott, C. J. and K. Matsuura, "Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance," *Climate Research*, Vol. 30, No. 1, 79–82, 2005.

[36] Chicco, D., M. J. Warrens, and G. Jurman, "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation," *PeerJ Computer Science*, Vol. 7, e623, 2021.

[37] Altman, D., D. Machin, T. Bryant, and M. Gardner, *Statistics with Confidence: Confidence Intervals and Statistical Guidelines*, John Wiley & Sons, 2013.

[38] Cumming, G. and S. Finch, "Inference by eye: Confidence intervals and how to read pictures of data," *American Psychologist*, Vol. 60, No. 2, 170–180, 2005.

[39] Crane, R. K., *Propagation Handbook for Wireless Communication System Design*, CRC Press, 2003.

[40] Hulst, H. C. and H. C. Van de Hulst, *Light Scattering by Small Particles*, Courier Corporation, 1981.

[41] Mie, G., "Beiträge zur optik trüber medien, speziell kolloidaler metallösungen," *Annalen Der Physik*, Vol. 330, No. 3, 377–445, 1908.

[42] ITU, "Specific attenuation model for rain for use in prediction methods," *Rec. ITU-R P.838*, 1992.