Sub-THz Sensing and Classification of Randomly Distributed Spherical Particles

Seyediman Nazerkakhki
Department of Electronic, Electrical and Systems Engineering
University of Birmingham
Birmingham, UK
sxn375@student.bham.ac.uk

James Elgy
Department of Electronic, Electrical and Systems Engineering
University of Birmingham
Birmingham, UK
j.elgy@bham.ac.uk

Alex Bystrov
Department of Electronic, Electrical and Systems Engineering
University of Birmingham
Birmingham, UK
a.bystrov@bham.ac.uk

Beth Jelfs
Department of Electronic, Electrical and Systems Engineering
University of Birmingham
Birmingham, UK
b.jelfs@bham.ac.uk

Marina Gashinova
Department of Electronic, Electrical and Systems Engineering
University of Birmingham
Birmingham, UK
m.s.gashinova@bham.ac.uk

Fatemeh Norouzian
Department of Electronic, Electrical and Systems Engineering
University of Birmingham
Birmingham, UK
f.norouzian@bham.ac.uk

***Abstract*— In many scientific and practical applications characterisation of randomly distributed particulates with different sizes and/or concentrations is fundamental to achieving reliable assessment of a given scene. This requires appropriate sensing methods using wavelengths relevant to the particle size, combined with suitable signal processing and machine learning (ML)-based classification techniques. In this paper, we investigate this problem using a large experimental dataset acquired through sub-THz radar in the 250–280 GHz range. The dataset consists of randomly distributed spherical particles with varying sizes and concentrations. We explore the ability of different ML algorithms: Extreme Gradient Boosting (XGBoost), Support Vector Machines (SVM), and Multilayer Perceptron (MLP), to classify either the number of particles or the size of the particles. The results indicate that XGBoost achieves the highest classification accuracy, above 90%, while incorporating higher numbers of features improves performance by up to 40%.**

Keywords*— *Feature Extraction, Machine learning, Performance Evaluation, Radar Measurements, Sub-THz

I. Introduction

Sub-THz radars, operating from 100 GHz to 1 THz, are now a practical reality for a multitude of applications, ranging from automotive radar [1] and meteorological monitoring [2] to security screening [3] and upcoming 6G infrastructure [4]. Due to their short wavelength and wide available bandwidth, radars operating in the sub-THz region provide an enhanced sensitivity to small objects, enabling an improved characterisation of media containing randomly distributed objects of diameters on scales of mm or smaller [5]. Recent advancements in machine learning (ML) have further enhanced this capability [6], with examples including weather radar [2], and pharmaceutical manufacturing [7].

This work investigates the classification of particle size and concentration of randomly distributed particles using sub-THz radar measurements. The frequency range of 250-280 GHz has been selected for this study due to a comparatively low atmospheric attenuation [8]. The dataset is acquired using a monostatic and quasi-monostatic radar configuration, consisting of spherical particles with variation in both diameter and concentration. Multiple ML algorithms – Support Vector Machines (SVM) [9], Multilayer Perceptron (MLP) [10], and Extreme Gradient Boosting (XGBoost) [11] – and feature combinations are compared in terms of their accuracy. The results indicate that XGBoost has the best performance due to its robustness against non-linear features arising from inter-particle coupling and an inbuilt regularisation to avoid overfitting [11]. Also, the results indicate that feature augmentation significantly increases classification accuracy across all models. This study demonstrates the feasibility of classifying the size and concentration of randomly distributed particles using physically meaningful features extracted from the sub-THz sensor into the ML algorithm.

II. Data Collection

Laboratory measurements were undertaken at the University of Birmingham THz measurement facility and consist of a series of step frequency continuous waveforms generated by a calibrated Keysight N5247B PNA-X vector network analyser (VNA). The frequency output of the VNA is extended by using a transmit-receive (Tx/Rx) Virginia Diode Inc (VDI) frequency extender up to frequency of 250-280 GHz and a receive-only (Rx) VDI extender [12].

The measurements include perfectly electrically conducting spherical scatterers with five different diameters ($D = 2$, 4, 6, 8, and 10 mm) with either 1, 2, 4, or 8 spheres in each measurement. The selected sphere sizes are slightly greater than the wavelength of the transmitted sub-THz wave ($\lambda \approx 1$ mm) resulting in an electrical size parameter ($\pi D/\lambda$) ranging from 6.3 at $D = 2$ mm and 31.4 at $D = 10$ mm. This implies that a geometric optics approximation to the scattering is appropriate for these measurements [13]. Consequently, the reflected signal is expected to be dominated by strong backscatter components and a small angular component [13], which can be respectively captured using co-polar monostatic and quasi-monostatic radar configurations.

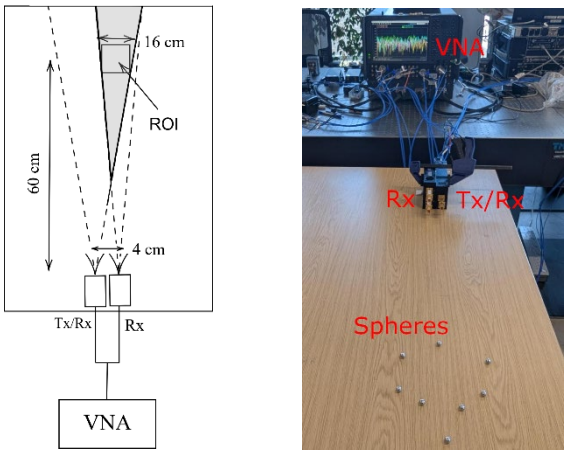
The radar parameters are detailed in Table 1, and the measurement topology is shown in Fig. 1. Spheres are placed on a wooden table within the intersecting region of both antenna beam patterns in a 16×16 cm region and offset from the radar by 60 cm. The table has been observed to produce minimal backscattered reflections, which are further reduced using background subtraction (discussed further in Sec. III).

Monostatic measurements are obtained using the Tx/Rx extender only, whereas the quasi-monostatic measurements are obtained using the Tx/Rx extender as a transmitter and the Rx extender as a receiver. The antennas are separated by 4 cm. This gives a bistatic angle to the spheres of $\approx 4°$, which we consider to be quasi-monostatic for these measurements. Measurements consist of both a background measurement and a sphere(s) measurement over the same region. To avoid unintentional bias in the training set due to particles' position,

measurements are repeated multiple times, with the sphere(s) being placed in different positions and arrangements for each measurement set while the antenna illumination area and setup remain consistent across the entire dataset. All sphere placements were restricted to the common illuminated region of the antenna beam patterns shown in Fig. 1. Throughout all measurement campaigns, the antenna configuration, table position, geometry, and radar-to-target separation remained unchanged.

Table 1 Sub-THz radar parameters.

Parameter	Value
Frequency	250-280 GHz
Power	2 dBm
Antenna Beam Angle	±10° (Tx/Rx) and ±5° (Rx)
Polarisation	Horizontal
Maximum Range	2 m
N. Sweeps	100



(a) (b)

Fig. 1. Measurement topology for 4-port measurements showing (a) schematic and (b) photograph. The Tx/Rx antenna and Rx only antenna and their respective beam patterns are shown.

III. Data Processing And Feature Extraction

A. Data Preprocessing

To suppress spectral leakage, a Hamming window is applied in the frequency domain. The signal is then converted to the range domain using an inverse Fourier transform, followed by range gating to isolate the region of interest (ROI), defined as the common measurement volume illuminated by all antennas. The ROI ($0.35 \leq r \leq 0.8$ m) is selected based on the known target location and is used to exclude additional clutter.

These steps are applied consistently to both measurement and background datasets. The processed signals are then used to compute the non-coherent background subtraction to suppress static clutter while preserving target reflections, given by:

$$P_{sig}(r) = 20\log_{10}\left(\left|R_{meas,i}(r)\right| - \left|\bar{R}_{bg}(r)\right|\right) \quad (1)$$

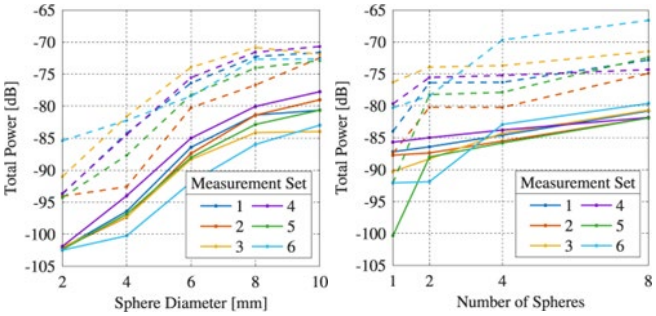
where $R_{meas,i}(r)$ represents the range profile magnitude at range r for the i-th measurement sweep, and $\bar{R}_{bg}(r)$ denotes mean background range profile magnitude, averaged over all background sweeps.

B. Feature Extraction

Three feature sets are considered in this work: 1) total received power, 2) the standard deviation (STD) of the received signal, and 3) phase represented using a polynomial best fit. The features are incorporated into the classification process cumulatively, where classification is first performed using the total reflected power, followed by the addition of the STD of the power within the ROI, and finally by including the phase as a third feature. The corresponding changes in classification accuracy due to this progressive feature augmentation are evaluated. However, the performance of classifiers based on individual features (e.g. power, STD or phase alone) or alternative feature combinations (e.g. pairwise subsets) is not investigated in this work and is left for future study. From (1), the total power is defined as:

$$T = \int_{0.35}^{0.8} P_{sig}(r)dr \quad (2)$$

The power is expected to be proportional to the number of illuminated spheres in the scene and, under a geometric optics approximation, $T \propto N\pi D^2/4$ where N is the number of spheres [13]. Therefore, T is expected to increase by 14 dB between the $D = 2$ mm and $D = 10$ mm, and by 9 dB when increasing the number of spheres from 1 to 8, giving a total possible variation of 23 dB. Fig. 2, illustrates the total received power with increasing diameter and concentration, showing a close match between measurement and prediction.



(a) (b)

Fig. 2. Quasi-monostatic (solid lines) and monostatic (dashed lines) total power for (a) 2 spheres with increasing diameter and (b) an increasing number of 6 mm spheres.

While T is easily computable and has an intuitive physical meaning, the dependence on N and D may lead to misclassification between a large number of small spheres and a small number of large spheres. To address this, the STD of P_{sig} is used to capture fluctuations in the scattering response. When spheres are added to the scene, additional peaks are introduced in P_{sig}. This feature provides additional discriminatory information, as higher variability in P_{sig} is expected in cases with a small number of larger scatterers, whereas lower variability corresponds to more uniform scattering environment, such as those consisting of a large number of small spheres.

Finally, phase information is also included as a feature. For a single sphere, the unwrapped measurement phase, ϕ, is expected to follow a linear trend with respect to frequency and is independent of sphere diameter. As additional spheres are introduced, inter-particle coupling and multiple reflections are expected to result in a non-linear phase behaviour with respect to frequency. To quantify this, a 4th order polynomial approximation is used [14]:

$$\phi(f) \approx a_4 f^4 + a_3 f^3 + a_2 f^2 + a_1 f + a_0 \quad (3)$$

where $a_4, a_3, \ldots, a_0$ represent the polynomial coefficients and f is frequency. These coefficients are used as input features,

with a_4, a_3, and a_2 expected to increase as the scattering environment becomes more complex.

IV. Results And Discussion

A. Classification Setup

The extracted features described in Sec. III-B are used as input to form feature vectors that are fed to the classification models. Specifically, three ML algorithms are investigated in this study: SVM, MLP, and XGBoost, using the Scikit-Learn Python library [15]. These models represent a range of approaches, including tree-based methods (XGBoost), neural network approaches (MLP), and kernel-based methods (SVM). SVM was selected as a strong baseline model, as it is widely used for classification of structured data and can handle non-linear decision boundaries using kernel functions [10]. MLP was chosen because neural networks have been popular in machine learning because they have the capability to address non-linear relationships [10]. XGBoost was chosen because it has consistently demonstrated high accuracy on structured and feature-based data, specifically when the classifier needs to learn the interactions and non-linear relationships among features [11].

The aim is to classify the sphere measurements based on both size and concentration, where both parameters are treated independently and two different classification tasks are performed. The dataset is partitioned into training, test, and validation sets with proportions of 70%, 20% and 10%, respectively.

Hyperparameter tuning was carried out using a 3-fold cross-validation grid search. Tuning was performed independently on both the size and concentration classification tasks. In both cases, we found that the optimal hyperparameter settings were very similar. Thus, for conciseness, we report only one set of hyperparameters (size) for each of the classifiers. The hyperparameters are summarised in Table 2.

Table 2 Specification of hyperparameters used in XGBoost, SVM, and MLP.

	Parameter	Value
XGBoost	Learning rate	0.1
	Max depth	5
	No. of estimators	600
	Min Child weight	1
	Subsample	0.8
	Colsample_bytree	1.0
SVM	Kernel	RBF
	Gamma	1
	Regularisation	10
	Scaling	StandardScaler
MLP	H. layers	64, 32
	Activation	tanh
	alpha	0.001
	Scaling	RobustScaler

B. Results and Discussion

The classification accuracy of all three classifiers and the effect of a richer feature representation are analysed in this section. The classification performance of all classifiers and features is summarised in Table 3. It is evident that the accuracy of all classifiers improves with a richer feature representation, with the lowest accuracy achieved when using only one feature (i.e. total power). Across all three classifiers, XGBoost provides the best performance for both size and concentration classification. An average accuracy of 94.4% and 82.7% is achieved for size and concentration classification, respectively, using all features with XGBoost.

To explain the improved performance of size classification, we note that, as shown in Fig. 2, changes in size result in a greater variation in T compared with increasing the number of spheres. For this reason, we should expect T to be a more important feature for size classification. To further analyse classification behaviour, Fig. 3 depicts confusion matrices using the best-performing classifier and feature set to gain more insight into classifier behaviour.

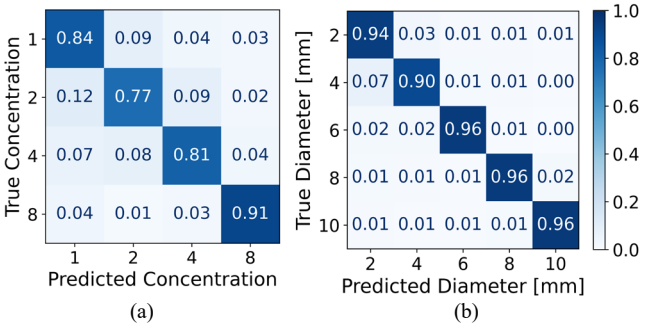


Fig. 3. Confusion matrix for XGBoost classifier with best-performing feature set (a) concentration and (b) size classification.

The classification results based on size are more diagonally dominant than those based on concentration, indicating that particles are more often classified into the correct size class as compared to the correct concentration class. Fig. 4 compares the classification accuracies across all classes individually for both concentration and size.

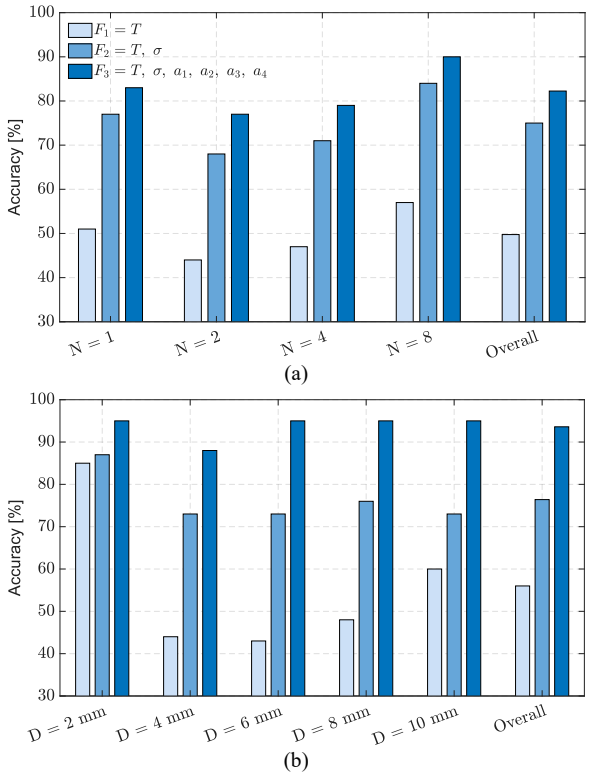


Fig. 4. Comparison of feature performance for XGBoost across different classes, showing (a) concentration classification and (b) size classification.

Table 3 performance comparison of different feature sets across all three classifiers.

Feature Sets		***T* only**			***T* and σ**			**T, σ, a_1, a_2, a_3, and a_4**		
Label	**Classifier**	**Overall Accuracy (%)**								
		Test	**Val.**	**Ave.**	**Test**	**Val.**	**Ave.**	**Test**	**Val.**	**Ave.**
Size	SVM	45.0	44.5	44.8	56.8	56.0	56.4	87.9	88.0	88.0
	XGBoost	56.1	54.4	**55.3**	76.2	75.7	**76.0**	94.3	94.55	**94.4**
	MLP	48.3	47.8	48.1	74.8	73.5	74.2	87.2	86.2	86.7
Concentration	SVM	36.0	37.1	36.6	46.3	47.8	47.1	75.6	76.2	75.9
	XGBoost	49.9	51.8	**50.9**	74.9	74.6	**74.8**	83.3	82.1	**82.7**
	MLP	44.3	45.7	45.0	57.6	57.8	57.7	77.4	77.9	77.7

The effectiveness of the classification is directly tied to the features employed. As shown in the figure, total power alone exhibits a limited classification ability, especially with regards to concentration. The addition of the STD boosts accuracy for all concentrations, with phase providing a more marginal increase. Size classification, on the other hand, shows a more significant contribution from the addition of phase as a feature, with a large increase in accuracy. Total power is more important for the $D = 2$ mm case, where power alone achieves an accuracy $> 80\%$. From Fig. 2a, we can observe that total power is tightly clustered for all measurement sets in the $D = 2$ case, whereas a greater variation is observed for larger spheres.

A key advantage of XGBoost is its aptitude for learning sophisticated non-linear relationships and interactions between the features, retaining its strong ability to generalise. On the other hand, SVM suffers from diminished performance, especially in the task of concentration classification, and MLP achieves intermediate performance without exceeding that of XGBoost.

V. Conclusion

A sub-THz sensing framework operating between 250 and 280 GHz was explored for classification of randomly distributed spherical targets with various diameters and concentrations. A calibrated dataset was collected, followed by the feature extraction based on total received power, magnitude change information, and polynomial coefficients computed from phase values.

Three classifiers, XGBoost, SVM, and MLP, were evaluated for their classification accuracy for differentiating spheres based on both size and concentration. Out of the tested algorithms, XGBoost was observed to perform best. The concentration classification consistently performed worse than the size classification task. We hypothesise that this is due to the quadratic growth in power as size is increased, resulting in a greater variation in total power and standard deviation compared with the linear increase due to introducing additional spheres. Finally, an increase in accuracy was observed, for all three models, when feature augmentation was performed. A systematic feature importance evaluation is planned for the near future.

Acknowledgment

This work has been supported by UK Research and Innovation (UKRI) under the project REPLENISH (10112182).

References

[1] F. Norouzian et al., “Rain attenuation at millimeter wave and low-THz frequencies,” IEEE Trans. Antennas Propag., vol. 68, no. 1, pp. 421–431, 2020.

[2] W. Schimmel et al., “Identifying cloud droplets beyond lidar attenuation from vertically pointing cloud radar observations using artificial neural networks,” Atmos. Meas. Tech., vol. 15, no. 18, pp. 5343–5366, 2022.

[3] R. Appleby and H. B. Wallace, “Standoff detection of weapons and contraband in the 100 GHz to 1 THz region,” IEEE Trans. Antennas Propag., vol. 55, no. 11, pp. 2944–2956, 2007.

[4] F. Taleb et al., “Propagation of THz radiation in air over a broad range of atmospheric temperature and humidity conditions,” Sci. Rep., vol. 13, no. 1, pp. 1–13, 2023.

[5] D. H. Choi, “Spectral scattering characteristics of subwavelength-sized spherical particles near Mie resonance modes probed by tightly focused terahertz waves,” Sci. Rep., vol. 15, no. 1, pp. 1-9, 2025.

[6] I. Cacciari and A. Ranfagni, “Unlocking terahertz technology with machine learning: A comprehensive review,” J. Appl. Phys., vol. 137, no. 22, pp. 1-17, 2025.

[7] Q. Yang et al., “Qualitative and quantitative analysis of caffeine in medicines by terahertz spectroscopy using machine learning method,” IEEE Access, vol. 9, pp. 140008–140021, 2021.

[8] G. A. Siles, J. M. Riera, and P. García-Del-Pino, “Atmospheric attenuation in wireless communication systems at millimeter and THz frequencies,” IEEE Antennas Propag. Mag., vol. 57, no. 1, pp. 48–61, 2015.

[9] R. Guido, S. Ferrisi, D. Lofaro, and D. Conforti, “An overview on the advancements of support vector machine models in healthcare applications: A review,” Information, vol. 15, no. 4, pp. 1-36, 2024.

[10] A. Chatterjee, J. Saha, and J. Mukherjee, “Clustering with multi-layered perceptron,” Pattern Recognit. Lett., vol. 155, pp. 92–99, 2022.

[11] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD), San Francisco, CA, USA, Aug. 2016, pp. 785–794.

[12] Virginia Diodes, Inc., “Vector network analyzer extenders,” [Online]. Available: https://vadiodes.com/vna-extenders-vnax/ (accessed Mar. 26, 2026).

[13] M. I. Skolnik, Radar Handbook, 3rd ed. New York, NY, USA: McGraw-Hill, 2008, pp. 14.1–14.16.

[14] S. Peleg and B. Portat, “Estimation and classification of polynomial-phase signals,” IEEE Trans. Inf. Theory, vol. 37, no. 2, pp. 422–430, 1991.

[15] F. Pedregosa et. al., “Scikit-Learn: Machine learning library in Python”, J. Mach. Learn Res., vol. 12, pp. 2825-2830, 2011