TURUN YLIOPISTO

Waste Material Classification with Microwave Reflection Data and Deep Learning

Data Analytics

Master's Degree Programme in Information and Communication Technology

Department of Computing, Faculty of Technology

Master of Science in Technology Thesis

Author:

Joona Helenius

Supervisor:

Doctoral Researcher Pouya Jafar Zadeh (University of Turku)

November 2025

Master of Science in Technology Thesis
Department of Computing, Faculty of Technology
University of Turku

Subject: Data Analytics
Programme: Master's Degree Programme in Information and Communication Technology
Author: Joona Helenius
Title: Waste Material Classification with Microwave Reflection Data and Deep Learning
Number of Pages: 68 pages, 10 appendix pages
Date: November 2025

This thesis examines the viability of microwave reflections as a data source for a Deep Learning -based classification model for the purposes of classifying waste materials. To perform this examination, a microwave reflection measurement system is created using a Vector Network Analyzer and two directional microwave antennae. Using this system a microwave reflection dataset consisting of three common recyclable material categories, plastic, metal and cardboard, is created by measuring the complex-valued S-parameter values of items from multiple angles over the frequency range 863 – 879 MHz. The size of the resulting dataset is 1112 samples split into 373 metal samples, 375 plastic samples and 364 cardboard samples. This dataset is used to train Multilayer Perceptron -networks in different data configurations to test the viability of microwave reflection data for material classification.

The results from different configurations indicate that microwave reflection data is suitable for classification of metal samples with the F1-score varying between 0.9149 and 0.9684. Differentiating between plastic and cardboard samples was much more unreliable with F1-scores of 0.7368 to 0.8505. To examine the full potential of this measurement system, research where a larger measurement bandwidth and more diverse set of material classes is required.

Keywords: Waste Management, Microwave, Deep Learning, Multilayer Perceptron, MLP

Diplomityö
Tietotekniikan laitos, Teknillinen tiedekunta
Turun yliopisto

Oppiaine: Data-analytiikka
Tutkinto-ohjelma: Tieto- ja viestintätekniikka
Tekijä: Joona Helenius
Otsikko: Roskamateriaalien luokittelu mikroaaltoheijastusdatan ja syväoppimisen avulla
Sivumäärä: 68 sivua, 10 liitesivua
Päivämäärä: Marraskuu 2025

Tämä tutkielma selvittää mikroaaltoheijastuksien toimivuutta datalähteenä syväoppimispohjaiselle luokittelumallille, kun tarkoituksena on luokitella jätemateriaaleja. Tätä selvitystä varten luodaan ensin mikroaaltoheijastuksia mittaava järjestelmä käyttäen vektoripiirianalysaattoria ja kahta suuntaavaa mikroaaltoantennia. Kyseistä järjestelmää käyttäen luodaan kolmesta yleisestä kierrätyskelpoisesta jätemateriaaliluokasta koostuva mikroaaltoheijastusdatajoukko. Käytetyt luokat ovat muovi, metalli ja pahvi. Datassa olevat mittaukset ovat kompleksiarvoisia S-parametrilukemia taajuusväliltä 863-879 MHz. Jokaista esinettä on mitattu useasta eri kulmasta, jotta saavutetaan kokonaisvaltaisempi profiili esineestä saatavista heijastuksista. Datajoukon lopullinen koko on 1112 näytettä, joista 373 on metallinäytteitä, 375 on muovinäytteitä ja 364 on pahvinäytteitä. Tällä datajoukolla koulutetaan monikerroksisia perseptroniverkkoja erilaisissa datan kokoonpanokonteksteissa, joiden suoriutumista käytetään mittaamaan mikroaaltoheijastuksen sopivuutta.

Eri kokoonpanoista saadut tulokset osoittavat, että mikroaaltoheijastusdata on käyttökelpoinen metallinäytteiden luokitteluun. F1-arvo metallinäytteille vaihteli kokoonpanosta riippuen arvojen 0.9149 ja 0.9684 välillä. Muovin ja pahvin erotteleminen sen sijaan oli huomattavasti epäluotettavampaa. F1-arvo muoville ja pahville vaihteli arvojen 0.7368 ja 0.8505 välillä kokoonpanosta riippuen. Jotta mittausjärjestelmän potentiaalia voitaisiin tarkastella kokonaisvaltaisemmin, tarvitaan tutkimusta, jossa olisi käytettävissä laajempi mittauskaista ja monipuolisempi joukko materiaaliluokkia.

Asiasanat: Jätehuolto, Mikroaalto, Syväoppiminen, Monikerroksinen perseptroniverkko, MLP

Table of Contents

1 Introduction

The yearly amount of waste generated globally is expected to increase from 2.01 billion tonnes in 2016 to 3.40 billion tonnes in 2050 due to urbanization, population growth and economic development and already modest estimates state that at least a third of this waste is already improperly handled [1]. In order to keep up with the rising amounts of waste, new technologies have to be developed to mitigate the environmental and health costs that are associated with improperly handled waste. The rise of artificial intelligence (AI) methods provides an excellent opportunity to create new technologies that could revolutionize the field of waste management, from assisting consumers with correct recycling practices to providing waste management facilities reliable waste sorting automation. Much of this potential is powered by AI's capabilities in handling highly multidimensional and noisy data [2], which are abundant in waste management.

The widespread use of AI is limited in part by the relative technological underdevelopment of waste management industry [2]. Many facilities are missing the required data collection procedures and tools to create high-quality data at required quantities to feed decision making pipelines [2]. It is thus no surprise that significant research effort is put into finding tools and methods that could reliably classify waste materials with the assistance of AI models. The most prevalent of these methods is the use of video or image data that is fed to AI for prediction, also known as computer vision. However, these applications have obvious issues with occlusion, bad lighting and other adverse conditions which can be prevalent in waste classification. Computer vision is also limited by regional differences in waste items appearance, making the creation of a high-quality and quantity dataset for model training extremely difficult.

This thesis aims to ease the task of solid waste classification by introducing an additional sensor input, microwave reflections, to be used in conjunction with other sensor data. Microwaves are a type of electromagnetic radiation with a wavelength of 1mm to 1m, corresponding to frequencies in the range 300 MHz to 300 GHz in the electromagnetic spectrum [3]. This range falls under the non-ionising radiation segment of the spectrum, meaning that microwaves are considered safe radiation and won't cause damage when coming in contact with living tissue [4]. They are used extensively in modern society, in wireless communications, medical applications, industrial heating, astronomy, microwave ovens and more [3]. By projecting a microwave signal to waste items and collecting the reflected signal characteristics, it could be possible to glean information about the items presented to a waste classification model without relying on the object's physical appearance.

To measure the properties of the reflected microwave signals, this thesis uses a vector network analyser (VNA), which is a device most commonly used in radio frequency (RF) engineering to test device properties and look for defective components. It operates by generating an RF signal to a device under test (DUT) and measuring the ratio of input and reflected signal power to determine the DUT's characteristics. Despite the fact that VNAs are not intended to be used as the signal generator component of a radar-like system, researchers in Thomas et al. [5] successfully classified cardboard objects covered in varying amounts of copper tape by using a VNA, microwave antennae and machine learning (ML) methods.

While the end goal would be to use microwave reflections as one data component in a sensor fusion system, this thesis will only test the performance of ML models using microwave reflection data as the only source of input data. This is done in order to isolate which types of materials can be classified using microwave reflections. The data will be used to train Multilayer Perceptrons (MLP), which are a type of Deep Learning (DL) model. The choice to use MLPs was made due a few primary factors: First, their simple implementation and superior performance over a support vector machine model in [5], which has been a source of significant inspiration for this thesis. Second, while the dataset used in this thesis is small for DL models, any future development will be performed with larger datasets and will benefit from DL methods. Third, if ever moved to a real-world context, non-DL methods will likely struggle to perform adequately [6].

1.1 Research Objectives and Research Questions

The two research questions posed in this thesis are as follows:

1. Can a vector network analyser be used to detect reflected microwave properties of plastic, cardboard and metal?

2. How well can an MLP network classify objects based on the object's reflected microwave properties?

Through these two research questions this thesis seeks to find whether or not microwave reflections contain enough information to be used as input data for waste material classification tasks. In order to answer these questions a dataset of microwave reflection measurements is created using a measurement system very similar to the one used by researchers in [5]. This dataset is created from household waste items belonging to one of three classes: plastic, cardboard and metal. These classes were chosen because they are common recyclable waste categories in Finland and are available in quantities that are reasonable for creating a ML dataset.

1.2 Previous Studies

Prior studies on computer vision -based waste classification are reviewed to illustrate current trends in the field. Each study faces different limitations, such as small datasets, overly idealized data quality and the absence of real-world challenges like occlusion or environmental disturbances. The microwave reflection experiment conducted by Thomas et al. is also presented for future reference.

In [7] researchers developed a convolutional neural network for classifying waste images from 12 different classes and a user-friendly web interface for uploading images for the network to label. The aim of the project was to ease the end-user strain of waste sorting by providing assistance with recycling practices. The best performing model in the paper was a pretrained VGG16-network that was transferred to a waste classification context using a dataset of roughly 15 500 images from an online repository of images that was relabelled to fit the purposes of the paper. It achieved a validation set accuracy of 77.62%.

In [8] researchers trained and evaluated four pre-existing models, ResNet50, InceptionV3, Xception and GoogleNet, on a four class dataset of cardboard, glass, metal and miscellaneous trash, with the aim of finding the best performing model for waste classification. The dataset consisted of 1 451 images of the aforementioned classes which were pre-processed from an existing dataset. The models received accuracy ratings of 90-100% with InceptionV3 performing at an accuracy rating of 100%.

In [9] researchers introduced a sensor fusion system of computer vision and weight to classify waste objects into recyclable or non-recyclable materials. 50 objects were used to create a dataset of 5000 training samples and 300 test samples for the purposes of training and evaluating a sensor-fusion -based model and a typical convolutional neural network for reference. The sensor fusion network attained a classification accuracy of 91.6% on the test set, surpassing the performance of the simple convolutional network by 11.6 percentage points.

In [5] researchers attempted to use microwave reflections combined with a ML model to differentiate between 5 known targets. The microwave reflection measurement system consisted of two wideband horn antennae as the microwave signal transmitter and receiver and a VNA to generate the incident signal and analyse the properties of the received signal. The objects were all differently shaped cardboard items which were either partially covered or fully covered by copper tape. During the measurement procedure each item was put on a turntable 230 cm away from the antennae and rotated 5 degrees at a time and measured 5 times from every angle for consistency.

After the item was fully rotated, the antennae were moved back 10 cm and the measurements were repeated. Each item was measured in this way for distances 230, 240 and 250 cm. The size of the total dataset was 5400 samples with each sample being 1001 equidistant S_{21}-parameter values over the frequency range of 5GHz-10GHz. The best in show ML model achieved a validation set classification accuracy of 96.9%.

1.3 Structure of the Thesis

Chapter 2 of this thesis will go over the necessary background information required for understanding the information contained in microwave reflections, the working principles of a VNA and DL model training and evaluation. Chapter 3 will outline the data measurement system, from measurement devices and software to the actual items included in the dataset and the data measurement principles. Chapter 4 will then explain the rationale for DL model hyperparameter optimization, model training and performance evaluation. Chapter 5 will delve further into the results of performance evaluation and finally chapter 6 will conclude the thesis with answers to research questions, suggestions for improvement and potential future work.

2 Background

2.1 Deep Learning

DL has gone through many different forms and names over its history which goes all the way back to the 1940's. The first developments toward DL occurred when a synthetic model of a neuron in the human brain was invented by McCulloch and Pitts [10]. This model would eventually be developed into the primary building block of artificial neural networks, which would later become known as DL. While the term DL is less prevalent in today's discussion of AI, modern AI technology in computer vision, large language models and more uses the fundamental structures of DL models as their core. This chapter will cover the fundamental concepts of deep neural networks and their training. Section 2.1.3 covers the most common performance metrics used to estimate ML model performance in classification contexts. While DL models can perform a multitude of different tasks, this thesis will only discuss DL in the context of classification.

2.1.1 Fundamentals

An intuitive way to describe DL is that it is a form of representation learning where the model attempts to learn more complex concepts sequentially from simpler concepts. In order to facilitate this concept-based learning the model has to develop different ways to represent the data by encoding it into different forms. This way of learning through representation is what allows DL networks to thrive when learning human-intuitive tasks like understanding images and sound, which have been very difficult for the ML models of the past. [6]

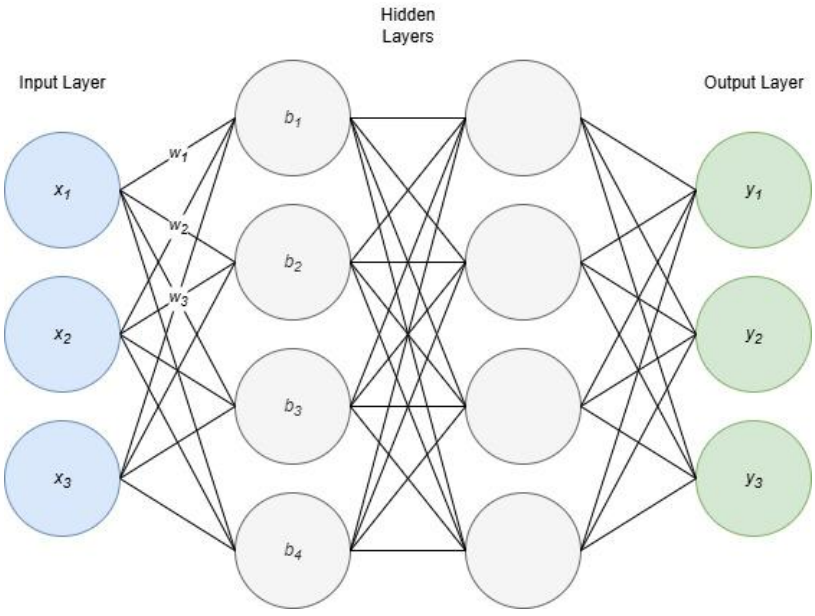

Figure 1. Structure of a simple DL network.

Figure 1 demonstrates the basic structure of a DL network. These DL models are also called MLPs. An MLP model consists of multiple perceptrons (nodes) arranged into multiple layers, categorized into input, output and hidden layers. The number of nodes in the input layer corresponds to the size of the input when flattened into a vector and the number of nodes in the output layer match the number of desired outputs. This is one node for binary classification and n nodes for an n-class classification. The number of nodes in the hidden layers is considered an optimizable hyperparameter and their role is extract features and encode the data into different representations.

Simple networks consist of a single input and output layer and at least two hidden layers to be considered a "deep" network [6]. Each node in each layer is connected to each other node in the following layer, forming a fully connected network. When data is fed to the network, it is passed from the input layer through each connection to the first hidden layer. Each connection changes the value of the input by multiplying it with some weight w. These weighted inputs are then summed in each node in the first hidden layer, and a bias value b is added. Mathematically speaking the value calculated at this step in a neuron for a network with n input values would correspond to

$$z = w_1x_1 + w_2x_2 + \cdots + w_nx_n + b. \quad (1)$$

This sum of values is then fed to a non-linear activation function. While many different activation functions have been developed and used in DL, the default choice has become the rectified linear unit (ReLU), which is defined as

$$ReLU(z) = \max\{0, z\} \text{ [6]} \quad (2)$$

The purpose of the activation function is to incorporate non-linearity to the model, which enables the model to solve tasks that are inherently non-linear. A toy example of these non-linear tasks is the exclusive-or problem [6]. Thus, the output of a node in an MLP network with a ReLU activation function is

$$a = ReLU(w_1x_1 + w_2x_2 + \cdots + w_nx_n + b) \quad (3)$$

This notation can be expanded to cover all the operations that occur when inputs are passed onto the first hidden layer

$$\boldsymbol{a} = ReLu(\boldsymbol{W}^T\boldsymbol{x} + \boldsymbol{b}), \quad (4)$$

where $\boldsymbol{W}^T$ is the matrix of weights that correspond to the connections between the input layer and the first hidden layer, $\boldsymbol{x}$ is the vector of inputs and $\boldsymbol{b}$ is the corresponding bias values in each node of

the hidden layer. In this case the ReLU activation is calculated elementwise for each individual node resulting in a vector of outputs $\boldsymbol{a}$.

The resulting vector of activations then becomes the new vector of input values for the next hidden layer and this process of multiplication, addition and taking the activation is repeated until the output layer is reached. The only change in the output layer is that the ReLU activation function is exchanged for a task-specific activation function, which is typically either the sigmoid activation function for binary classification tasks and the SoftMax activation function for multiclass tasks. The output of a DL model is called a prediction which in the classification case is a floating point value corresponding to what the model considers is the probability of the input belonging to the predicted class. As such, it has a maximum value of 1 and a minimum value of 0. In the binary case the model outputs a single value as its prediction and in the multiclass case it outputs a vector of length n where n is the total number of predictable classes.

2.1.2 Model Training

Training DL networks involves repeatedly applying two processes: the forward pass and the backward pass. The forward pass is the passing of an input through the network from the input layer to the output layer such that the model makes a prediction on the input. The mathematics of the forward pass are essentially a repeated application of formula 4, with the exception that the output layer uses a different activation function. The prediction is then compared with the true value to calculate the amount of error made by the network. The typical way to process this is to assign a cost, usually called a loss, to each prediction such that correct predictions have low or 0 loss and incorrect predictions have a high loss. The most common loss function used in classification is the cross-entropy loss, also known as log-loss. The following formula is for binary cross-entropy (BCE), which is used in binary classification

$$BCE = -\frac{1}{N}\sum_{i=1}^{N}(y_i \log(p_i) + (1-y_i)\log(1-p_i)), \quad (5)$$

where N is the number of samples, y_i is the true label of the i:th observation and p_i is the predicted probability of the i:th observation belonging to class 1. It is noteworthy that due to the properties of logarithms, values closer to 0 are given large negative values whereas values close to 1 trend towards 0. Thus, BCE gives a very large cost to highly incorrect predictions ($y_i log(0.001) = -3$, when $y_i = 1$) and no cost to perfectly correct predictions ($y_i log(1) = 0$, when $y_i = 1$).

BCE is incompatible with multiclass classification due to its reliance on a singular binary value for the correct predicted class. For this reason, a multiclass variant of BCE is used that can calculate the loss for a one-hot encoded vector of prediction probabilities: categorical cross-entropy (CCE). Presented mathematically it is

$$CCE = -\frac{1}{N}\sum_{i=1}^{N}\left(\sum_{j=1}^{C}\left(y_j \log p_j\right)\right), \tag{6}$$

where C is the total number of classes being predicted for, y_j is the true label of the j:th element in the vector of true classification labels and p_j is the predicted probability of class j being true. While CCE is visually different from BCE, in practice it functions nearly identically.

The backward pass is where the model "learns" from the mistakes made in the forward pass. On an intuitive level, the mistakes that the model made in the forward pass are passed back through the network and used to teach the network to avoid similar mistakes in the future. Concretely, this learning happens through updating the weight and bias parameters in the network. The way these parameters are updated is governed by two things: backpropagation and the optimizer. Backpropagation is an algorithm that is used to efficiently calculate the partial derivatives of the loss for each and every parameter in the network using the chain rule of derivatives.

$$\frac{dE}{dw_{ij}} = \frac{dE}{da_j}\frac{da_j}{dz_j}\frac{dz_j}{dw_{ij}} \tag{7}$$

To calculate the partial derivative of the error with regards to w_{ij}, which is the i:th weight for neuron j, the chain rule has to be applied twice. Here z_j is the pre-activation sum of weighted inputs and a_j is the post-activation output value of the neuron. The right-most fraction can be greatly simplified by using the properties of neural networks. The change in pre-activation sum of values with regards to w_{ij} is only ever dependent on the input x_i to that specific weight.

$$\frac{dz_j}{dw_{ij}} = \frac{d}{dw_{ij}}\sum \boldsymbol{w}_j\boldsymbol{x} = \frac{d}{dw_{ij}} w_{ij}x_i = x_i \tag{8}$$

These partial derivatives calculated using backpropagation form gradients of the loss function. These gradients can be used in gradient-descent -based optimization algorithms to update model parameters to train the network. In the simplest terms the updating of the network parameters is done by multiplying the partial derivative by a learning rate $\eta < 1$ (typically 0.01 or smaller) and changing the value of the corresponding parameter by that amount. The role of the learning rate is to reduce the amount each parameter is updated to avoid large oscillations in these updates. These can cause the learning process to overshoot parameter updates from the gradient minima and slow

down convergence. This simple update strategy is called Stochastic Gradient Descent (SGD), presented in formula 9 and it functions as the basis for many other optimizers. [6]

$$w_t = w_{t-1} - \eta \frac{dE}{dw_{ij}} \tag{9}$$

The term w_{t-1} corresponds to the value of w_{ij} at time step $t-1$ and w_t is the new value of w_{ij} after the parameter update.

For this thesis two more developed optimization algorithms are presented which adaptively adjust the learning rate for each individual parameter in the network: RMSprop [11] and Adam [12]. RMSprop updates the naive SGD algorithm by adding a term which calculates an exponentially weighted moving average of previous parameter updates [6]. It is especially effective in cases where the gradient function is convex in nature [6].

$$v(w,t) = \gamma v(w,t-1) + (1-\gamma)\left(\frac{dE}{dw_{ij}}\right)^2 \tag{10}$$

$$w_t = w_{t-1} - \frac{\eta}{\sqrt{v(w,t)}}\frac{dE}{dw_{ij}} \tag{11}$$

The term γ is the forgetting factor, usually initialized as 0.9 [13].

Adam can be considered to be a combination of RMSprop and momentum. It is typically extremely robust to hyperparameter selection [6]. Its implementation of momentum is very similar to how RMSprop calculates weighted moving averages.

$$m(w,t) = \beta_1 m(w,t-1) + (1-\beta_1)\frac{dE}{dw_{ij}} \tag{12}$$

$$v(w,t) = \beta_2 v(w,t-1) + (1-\beta_2)\left(\frac{dE}{dw_{ij}}\right)^2 \tag{13}$$

$$\hat{\mathrm{m}}(w,t) = \frac{m(w,t)}{1-\beta_1^t} \tag{14}$$

$$\hat{\mathrm{v}}(w,t) = \frac{v(w,t)}{1-\beta_2^t} \tag{15}$$

$$w_t = w_{t-1} - \eta \frac{\hat{\mathrm{m}}(w,t)}{\sqrt{\hat{\mathrm{v}}(w,t)}+\epsilon} \tag{16}$$

Terms β_1 and β_2 are forgetting factors which are usually initialized as 0.9 and 0.999, respectively, β_1^t and β_2^t are bias correction terms that denote β_1 and β_2 to the power of t and ϵ is a small scalar (10^{-8}) used to avoid divisions by zero. [12]

The central goal of model training is for the model to be able to predict accurately on previously unseen samples. This trait is referred to as generalization. The samples that form the dataset used to train the model are examples drawn from some unknown probability distribution. In order for the model to correctly predict unseen samples, it has to learn this distribution using the training data. If the model learns the distribution of the training data rather than the underlying probability distribution, the model loses its ability to generalize. This phenomenon is called overfitting. To follow the progress of model training and avoid overfitting, ML model training uses multiple mutually exclusive datasets such that only one dataset is used for training and other sets are used to validate and test the model's ability to generalize. A telltale sign of overfitting is when model performance on the training set is very high, but performance on the validation and test sets are low. [6]

The use of multiple sets of data is at odds with ML methods' reliance on large quantities of training data to be able to reliably generalize. As these separate sets are typically generated by splitting an initial dataset into multiple pieces, there is a clear contradiction between the goal of maximizing training data and ensuring that models don't overfit. Set splitting is also very commonly a random process that can cause some sections of the true probability distribution to be underrepresented in the training data. To evaluate the performance of ML models more reliably and reduce the impact of randomness, cross-validation is typically used. [6]

Cross-validation is a method where multiple versions of the same model are trained on differing configurations of training and validation data. While many different versions of cross-validation exist, for the purpose of thesis only K-fold cross-validation is presented for brevity. In K-fold cross-validation the training data is split into *k*-folds which are partitions of the data of equal size. The model is trained *k* times with each fold serving as the validating set once and the rest of the data is used for training. By aggregating the performance of the model across the different folds, the reliability of the performance estimation can be increased. [6] Figure 2 shows the principle of how data would be split and arranged for a single fold in 5-fold cross-validation.

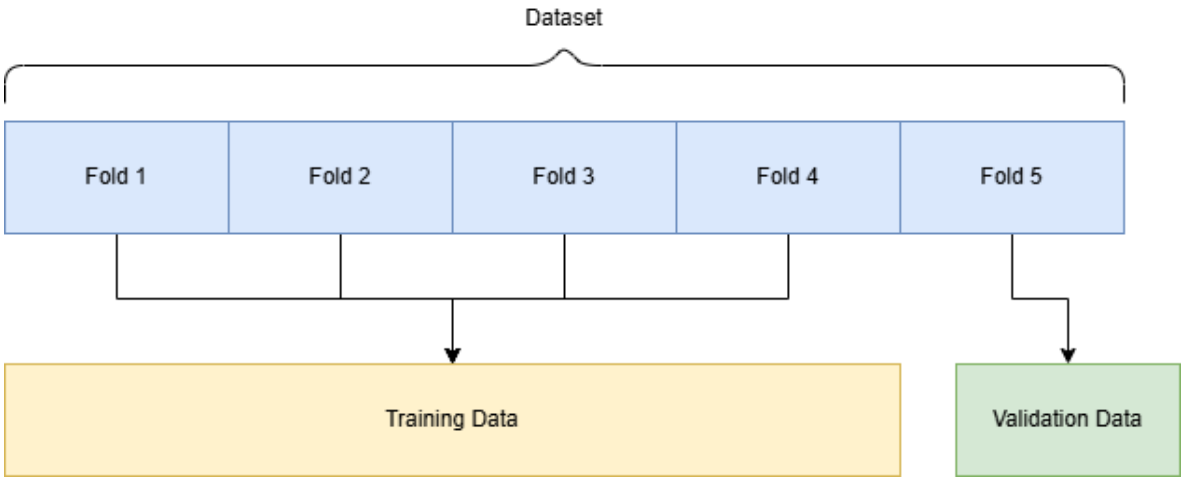

Figure 2. Example of data splitting in 5-fold cross-validation and arrangement of data into one of the five training configurations.

2.1.3 Machine learning performance metrics

A confusion matrix is a commonly used ML performance visualisation tool. By showing the relationships between true and predicted classes, a confusion matrix is able to clearly visualize where the classifier is making mistakes and where it is performing well. The confusion matrix is also the base for many performance metrics. The structure of a binary case classification matrix is shown in figure 3. [14]

Class	True 1	True 0
Predicted 1	True Positive	False Positive
Predicted 0	False Negative	True Negative

Figure 3. Binary confusion matrix and prediction types.

The four types of predictions that a classifier can make are also present in figure 3. These are the True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN). [14]

Accuracy is simply defined as the number of correctly predicted samples divided by the number of total samples. In a binary classification case it can be defined mathematically as:

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (17)$$

In a multiclass classification case accuracy can be directly calculated from a confusion matrix by taking the sum of the matrix diagonal and dividing by the total number of predictions. [14] Accuracy is a great baseline tool in assessing the overall performance of a predictor but it has shortcomings, especially in cases where significant class imbalance exists in data. To obtain a more robust evaluation of a model's performance, additional performance metrics are used.

Precision measures how many of the cases predicted positive by the model are actually positive. In a binary classification case precision can be defined mathematically as follows

$$Precision = \frac{TP}{TP+FP} \quad (18)$$

Therefore, precision measures how trustworthy the positive labels assigned by the model are. [14]

Recall is the measure of how many truly positive samples were predicted as positive. The mathematical definition for recall in the binary case is as follows:

$$Recall = \frac{TP}{TP+FN} \quad (19)$$

The reason to measure recall is to assess how well the model finds all of the positive cases in the set. [14]

And last is the F1-score which is the harmonic mean of precision and recall. The binary F1-score is defined as follows:

$$F1 = 2 * \left(\frac{Precision*Recall}{Precision+Recall}\right) \quad (20)$$

As a harmonic mean of two values which themselves have values between 0 and 1, the F1-score is also a real value between 0 and 1 with 1 being best. It rewards models which attain balanced values for precision and recall over very unbalanced precision-recall combinations. [14]

Class	True 1	True 2	True 3
Predicted 1	TP	FP	FP
Predicted 2	FN	TN	TN
Predicted 3	FN	TN	TN

Figure 4. Multiclass classification matrix and classification types with regards to class 1.

Class	True 1	True 2	True 3
Predicted 1	TN	FN	TN
Predicted 2	FP	TP	FP
Predicted 3	TN	FN	TN

Figure 5. Multiclass classification matrix and classification types with regards to class 2.

Due to precision, recall and F1-score being based on a binary classification context, they are not applicable as is to a multiclass classification context as prediction types are no longer static. This is highlighted in figures 4 and 5 which demonstrate the prediction types in cases where class 1 and class 2 are being examined respectively. Therefore, the formulae for these metrics have to be extended slightly. In a multiclass classification context precision and recall are calculated individually for each class and averaged out in one of two ways: micro or macro [14]. In the micro-average no attention is given to class distributions, and everything is evaluated as a lump [14]. This has the side effect of making the micro F1-score reduce down to being equivalent to accuracy and for this reason the micro case is not elaborated further. All multiclass precisions and recalls in this

work will be calculated as macro averages by default. In the macro case precision and recall are first solved for each class individually:

$$Precision_k = \frac{TP_k}{TP_k + FP_k} \tag{21}$$

$$Recall_k = \frac{TP_k}{TP_k + TN_k} \tag{22}$$

For average precision and recall the simple mean of each class-wise result is taken. This way of solving the means also ignores any class-imbalances present in the set as the performances for each class are weighted equally. [14]

$$MacroAveragePrecision = \frac{\sum_{n=1}^{k} Precision_n}{k} \tag{23}$$

$$MacroAverageRecall = \frac{\sum_{n=1}^{k} Recall_n}{k} \tag{24}$$

Finally, solving the macro F1-score is attained by solving for F1-score using average precision and recall:

$$MacroF1 = 2 * \left(\frac{MacroAveragePrecision * MacroAverageRecall}{MacroAveragePrecision + MacroAverageRecall}\right) \tag{25}$$

Due to the macro F1-score being the harmonic mean of the class-wise average precision and recall, a high score implies that the classifier has stable performance across each class.

2.2 Microwaves

Microwave propagation is the phenomenon of microwaves travelling through a medium. In an idealized context where microwaves aren't subject to any forces that affect their propagation, they travel in a straight line without any loss in signal strength. This propagation context is called free space. As was alluded to, free space is an idealized medium and in real-world conditions microwaves are subjected to multiple different phenomena that can affect signal direction and strength. [3]

2.2.1 Microwave Behaviour and Propagation Medium Change

For this experiment the most relevant phenomena are ones where microwaves collide with an object of different propagation medium. These phenomena —reflection, refraction, transmission, and repolarisation —contain information about the material and its immediate environment. For any

collision one or more of these events can occur simultaneously [15]. The specifics of which interactions happen and to what extent are governed by the dielectric properties of materials and how the space occupied by the material is altered by the material's electrical and magnetic properties, mainly electrical permittivity and magnetic permeability. These parameters are measures of the amount of charge needed to generate one unit of flux in a medium and of the material's ability to support the formation of a magnetic field inside itself respectively. These dielectric properties in most materials are directly related to its molecular composition and are frequency-dependent. [4]

Figure 6 illustrates a situation in which microwave radiation passes from one propagation medium into another. In this case microwave radiation is simultaneously reflected, refracted, attenuated, and transmitted due to the change in this medium. Transmission is the event where microwaves pass through objects. Reflection on the other hand causes microwaves to bounce off of the surface of medium two back to medium one at an angle that is governed by Snell's and Fresnel's laws of reflection. Refraction causes the redirection of the transmitted wave in medium two. This is caused by a change in the propagation speed of the wave in medium two. [15]

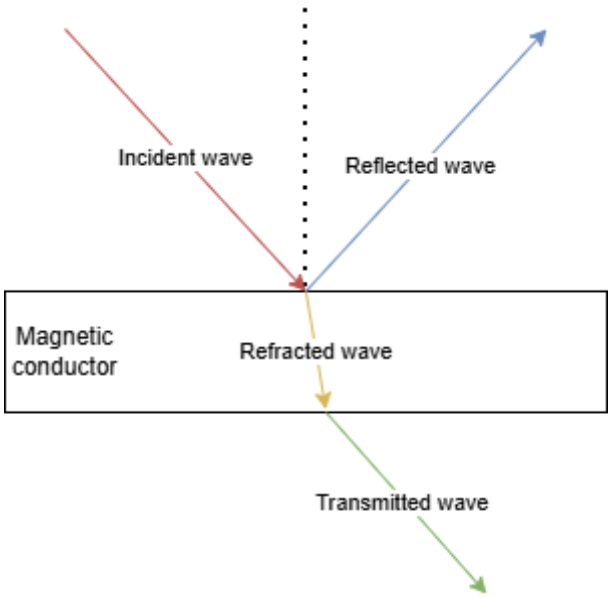

Figure 6. Microwave collision with a magnetic conductor. [15]

If microwave propagation were lossless, the energy of the propagated signal would not diminish over any distance. However, in real-world scenarios these interactions are always lossy. Interactions where signals lose energy while propagating are collectively called attenuation and can be caused by, for example, absorption and dispersion. In air attenuation is usually negligible over small distances and low frequencies but becomes a problem at frequencies higher than 10 GHz or high

atmospheric moisture conditions due to atmospheric gases such as oxygen and water vapour absorbing microwave radiation [3]. When colliding with obstacles, microwaves are also attenuated. This is mainly caused by losses during transmission. While passing through an object, energy is absorbed by the object and converted into heat, reducing the remaining signal power. If a material is a very strong absorber of a frequency of electromagnetic radiation, it will also tend to have strong surface reflective properties [16]. [15]

2.2.2 Echo Power Density and Radar Cross Section

The measurement system used in this thesis is in many ways similar to a RADAR (RAdio Detection And Ranging) -system. Both systems obtain information about objects using echo signals and for that reason it is necessary to explain a concept that is key to how echo signals behave in RADAR-systems. This concept is the Radar Cross Section (RCS) and it partially governs the power density of the echo signal [3].

The simplest implementation of a RADAR consists of four components: a signal transmitter, a signal receiver, a transmitting antenna and a receiving antenna. The main use case of a RADAR system is to locate and track the movement of distant objects using RF waves. By transmitting high-powered radiation towards a target, receiving the echo signals that the target reflects back and measuring the time difference between signal transmission and retrieval, the RADAR is capable of measuring the distance between the RADAR and the target. Repeated signal pulses allow for more information to be collected and, thus, the target's direction and velocity can also be determined. [3]

When the incident radiation hits the target, the radiation is dispersed and re-radiated in multiple directions. Only the part of the radiation which is re-radiated back toward the RADAR system is intercepted by the receiver. For typical RADAR implementations where distance is measured, Echo Power Density (EPD) provides the mathematical grounds for calculating the maximum distance of detection for a given RADAR system. For the system proposed in this thesis EPD provides the mathematics of the power density intercepted in the receiving antenna

$$EPD = \frac{P_t G_t \sigma}{(4\pi R^2)^2}\ W/m^2, \tag{26}$$

where P_t is the power radiated by the transmitting antenna, G_t is the directional gain of the transmitting antenna and R is the distance to the target. σ is the RCS of the target, the value of which is dependent on the properties of the target and the incident radiation. Such properties include things like the target's size and type, but also frequency, polarization, and the relative incident and

reflective angles with regards to the target. Even small changes in the incident angle can cause the value of the RCS to shift dramatically, which causes fluctuations of the RCS when attempting to use a RADAR on moving targets. [3]

2.3 Vector Network Analyzer

Microwave frequencies make taking simple electronic measurements impossible without specialized equipment. This issue is caused by the parasitic elements of typical electrical circuits becoming too prominent at higher frequencies for reliable measurements to be taken [3]. In order to be able to measure microwave signal properties this thesis uses a VNA as the measurement device. A VNA is a combined signal generator and receiver component such that it measures the operating properties of the DUT under a wide range of frequencies.

The DUT can be categorized as a single-, two- or multi-port network. A single-port network only contains a single input port. The VNA would in this case measure the amplitude and the phase of the incident signal going into the DUT, known as V^+, and the reflected signal coming back into the input port from the DUT, known as V^-. The ratio between the reflected signal and the incident signal is called the reflection coefficient Γ and is defined mathematically in equation 24. [17] An example of this type of device would be an RF antenna.

$$\Gamma = \frac{V^-}{V^+} \tag{27}$$

A two-port network instead contains a single input port and a single output port. In two-port networks the properties of the DUT can be described by a 2-by-2 matrix of coefficients called S-parameters, also known as scattering parameters. Each parameter S_{nm} describes the ratio of the signal output from port n divided by the signal input from port m when no other signal inputs are present in the DUT. The matrix and general equation of S-parameters are shown in equations 25 and 26. Each S-parameter is also known by a more colloquial name. S_{11} and S_{22} parameters are known as the input and output reflection coefficients respectively, as they describe the ratio of reflected signal at the input and output ports. The S_{21} parameter is called the forward transmission coefficient as it describes the ratio of the signal going through the device from the input port to the output port. The S_{12} is called the reverse transmission coefficient following a similar logic. It conveys the ratio of the signal going through the device from output to input. [17]

$$\begin{bmatrix} V_1^- \\ V_2^- \end{bmatrix} = \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix} \begin{bmatrix} V_1^+ \\ V_2^+ \end{bmatrix} \tag{28}$$

$$S_{nm} = \frac{V_n^-}{V_m^+}\Big|_{V_{m\neq n}^+=0} \quad (29)$$

These responses given by DUTs in RF engineering are dependent on the frequency of the input signal. Therefore, VNAs are capable of outputting a wide range of signal frequencies, from a few kilohertz to hundreds of gigahertz [18][19]. The specific range that a single VNA is able to produce is device-specific, but examples of high range equipment that can output between the range of 70 kHz and 220 GHz exist [19]. These high-end devices are also typically very expensive and mainly used in industrial applications due to their prohibitive cost. However, in addition to the costly high-end devices, handheld versions of VNAs are available. These devices offer much more modest capabilities but at significantly lower costs and have opened vector network analysis for a larger audience of users. The VNA used in this experiment is specifically a handheld VNA as the cost of an industrial-grade VNA would be prohibitively for this thesis.

Both types of devices are prone to three kinds of errors during measurement: systematic errors, random errors and drift errors. The effect of systematic and drift errors can be reduced by calibrating the device before use. This is done by applying a combination of calibration standards whenever the measured frequency range is changed. These calibration standards correct measurement errors mathematically during measurements allowing VNA devices to give much more accurate results. These calibrations are not perfect and some amount of error will persist even after calibration. [17] There exist many different calibration standards and standard combinations that can be used for VNA calibration but for the purposes of this work only the traditional two-port SOLT calibration combination is explained.

SOLT is an acronym of Short, Open, Load and Through, the four calibration standards that it comprises of. The SOLT calibration process in the nanoVNA-f V2 device [20], which is the device used in this thesis, is composed of the following steps:

1. Configure the device’s measurement frequency span.
2. Apply the short calibration component to the Tx port of the VNA, then calibrate the device’s short parameter.
3. Apply the open calibration component to the VNA's Tx port and calibrate the device's open parameter.
4. Apply the load calibration component to the Tx port of the VNA and calibrate the device’s load parameter.

5. Connect the Tx port to the Rx port with a coaxial cable, then calibrate the through parameter of the device.

Each of these four calibration components induces a different type of standard electrical parameter to the VNA which can be used to solve the 7-term error box model of a two-port system [17]. An oversimplified explanation of the different calibration standards is as follows: the Open standard models a delay from the connector to a small parasitic capacitor, the sort is a delay from the connector to a small parasitic inductor, the through is simply a delay and the load is a 50 (ohm) impedance, also called a perfect impedance [21].

3 Data

In order to train a DL based ML model and test the performance of the proposed material classification solution, a dataset must be obtained. Due to not finding any publicly available datasets that could be used for this experiment a dataset was created using low cost and existing equipment available at the University of Turku and household trash items. The core of the data collection setup follows the procedure demonstrated by [5] in their microwave reflection experiment with some adjustments to align the process to a materials classification task and account for the lower available budget. In essence, the proposed measurement system will attempt to measure the complex-valued EPD intercepted by the receiving antenna at a range of frequencies. In this chapter the dataset creation process and the resulting dataset are explored. This includes specifying the equipment that were used in the measurement of samples, specifying the data measurement process, outlining the resulting dataset and performing data exploration to find insights from the data.

3.1 Equipment

3.1.1 Antennae

This experiment uses the same core measurement equipment as [5], those being two antennae and a VNA. However, for both devices a more affordable version has been obtained due to budgetary constraints. Both antennae are Laird Technologies S8655PR circular polarity RFID panel antennae [22]. The main reason these antennae were chosen over a competing pair of antennae was their directional nature. Directionality means the antennae can be positioned so that less of the transmitted radiation is directly intercepted by the receiving antenna. This is assumed to reduce the noise in the receiver antenna and therefore increase the capability of the system to measure the reflected component of the signal.

In the datasheet for the S8655PR antennae provided by the manufacturer, the operating frequency range of the antenna is stated to be 865 – 870 MHz. To see if these antennae could be used outside of the provided operating frequencies, the devices were tested using the nanoVNA-F V2 device outlined in the next section of this chapter. It was discovered that both antennae retain competitive S_{11} parameter values over the frequency range of 863 MHz – 879 MHz. This frequency range was then chosen as the measurement span for two reasons: the competitive S_{11} parameter values that the antennae demonstrate and the frequency band not being in use by Finnish telecommunications as outlined by the Finnish Transport and Communications Agency TRAFICOM [23].

The minimum and maximum S_{11} values for each antenna were -19.57 dB and -17.13 dB for antenna 1 and -18.46 dB and -17.32 dB for antenna 2 over the measurement frequency range. The differences in measured S_{11} values between the antennae are likely due to small manufacturing differences or noise in the measurement environment or the VNA. The general guideline for antenna quality is that an S_{11} parameter value of less than -10 dB is considered an acceptable value and anything lower is considered an improvement [24]. An S_{11} of -10 dB means that in practice 90% of the input power is effectively converted by the antenna into radiation and only 10% is reflected back into the transmission line. Using this analysis it was deemed that the antennae are capable outside of the frequency range provided by the manufacturer as the effective power conversion is roughly 98-99% over the chosen measurement frequency range. The full S_{11} values for both antennae over the measurement frequency range are shown in figure 7.

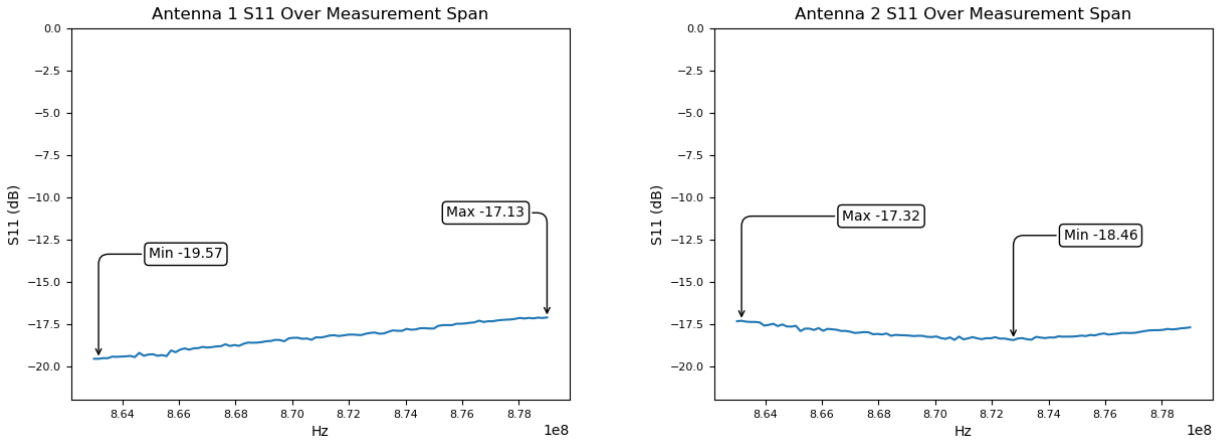

Figure 7. Left: antenna 1 S11 parameter values over the measurement frequency range. Right: Antenna 2 S11 parameter values over the measurement frequency range.

3.1.2 nanoVNA-F V2

The VNA in this experiment is a nanoVNA-F V2 by Sysjoint [25], which a small portable VNA with a minimum and maximum operating frequency of 50 kHz and 3 GHz respectively. The sweep frequency span can be freely calibrated to any frequency range within the minimum and maximum constraints and the device will measure 101 evenly spaced datapoints during each sweep. The dynamic range of the VNA is 50 dB for the S_{11} parameter when measuring under the 1.5 GHz frequency threshold. For the S_{21} this dynamic range value increases to 70 dB under the same conditions. This means that the difference between the largest and the smallest value that the VNA can measure in each sweep is 50 dB for S_{11} and 70 dB for the S_{21}. The VNA was calibrated with a

standard SOLT calibration (Short, Open, Load, Through) over the frequency range of 863 MHz and 879 MHz.

There are many drawbacks to using a nanoVNA as opposed to a regular VNA such as a smaller operating frequency range, a smaller number of datapoints measured per sweep, increased noise in measured values, especially at higher frequencies, and a smaller dynamic range. In this experiment some of these drawbacks are less severe due to the nature of the measurement setup. The operating frequency of the nanoVNA-F V2 is perfectly suitable for the used antennae as was shown previously. Due to the very limited operating frequencies of the antennae the problem caused by the small number of measured datapoints diminishes as well. As the planned sweep spans over 16 MHz with 101 datapoints per sweep, the distance between two consecutive datapoints is only $\frac{16\ MHz}{100} = 0{,}16$ MHz. This will likely avoid problems that would arise from an overly sparse data resolution like loss of detail in the data.

The two hardware related problems that cannot be directly dismissed are the ones relating to dynamic range and noise. A lower dynamic range might cause the nanoVNA to miss reflected signals that are too low in signal strength for the device to measure. With higher grade VNAs, these signals might have been detectable. In order to address the problems that could arise from the lower dynamic range the decision was made to perform only close-range measurements of the objects. As the power density of the electromagnetic radiation decreases by a factor of distance squared (see formula 23), closer range measurements will mean that the electromagnetic field strength will be larger and, thus, more likely to be detectable by the receiving antenna.

Noise can cause the measurements taken to be more unreliable due to variations in measured values that would not be present with higher grade VNAs. The presence of random measurement noise can also cause randomly occurring trends that create false bases for prediction. Observations regarding the presence and significance of noise in the data are performed in the data exploration segment of this chapter and later on in the concluding chapter of this thesis.

3.1.3 Other Equipment

Additional relevant equipment during this experiment is the nanoVNASaver [26] application and the measured items. nanoVNASaver is a free software that can be used for displaying visualisations of nanoVNA measurements and creating touchstone files of sweep data. Touchstone files are a file format for saving frequency dependent S-parameter data in ASCII [27]. Additionally nanoVNASaver can be used to increase the number of datapoints per sweep by splitting frequency

bands into multiple segments and performing a sweep over each segment. nanoVNASaver is licensed under version 3 of the GNU General Public License [28]. No changes were made to the nanoVNASaver application in this experiment.

The measured objects are primarily Finnish household trash items and some metal utensils. Three classes of objects were chosen from common item material types: metal, plastic and cardboard. These specific classes were chosen due to their ease of availability and all three being common types of recyclable waste in Finland. The primary objective when choosing items for the dataset was to obtain a maximally diverse dataset by selecting items that are either different in shape or represents a different material sub-type (for example an aluminium drink can vs a tin). In some cases this required deformation of the original object shape. A large portion of the metal objects were cylindrical in shape because the available metal waste was relatively homogenous. Therefore, many of the metal objects in this experiment were deformed in some way. The same was done to a few objects in the cardboard item set but due to the diversity of shapes in the plastic items, no deformation was required for that set. Examples of metal items and deformation can be found in figure 8.

Figure 8. Examples of metal objects and deformation. Left object has no deformation and the two objects to the right show different styles of deformation.

3.2 Data Gathering

Typically a two-port VNA measurement measures the properties of a single device that is connected via cables to both the transmitting and receiving ports of the VNA. In this experiment the two-port measurement is altered such that one antenna is connected to the transmitting port and one antenna is connected to the receiving port. Thus, the signal from the transmitting port is converted to microwave radiation at the transmitting antenna. This radiation is then directed towards an object and when the incident radiation collides with the object a combination of reflection, absorption and transmission will occur. Some of the reflected signal will be directed towards the receiver antenna which will capture this microwave radiation and convert it back into a signal such that the VNA can compare the properties of the transmitted and received signals. An image of the measurement system is shown in figure 9.

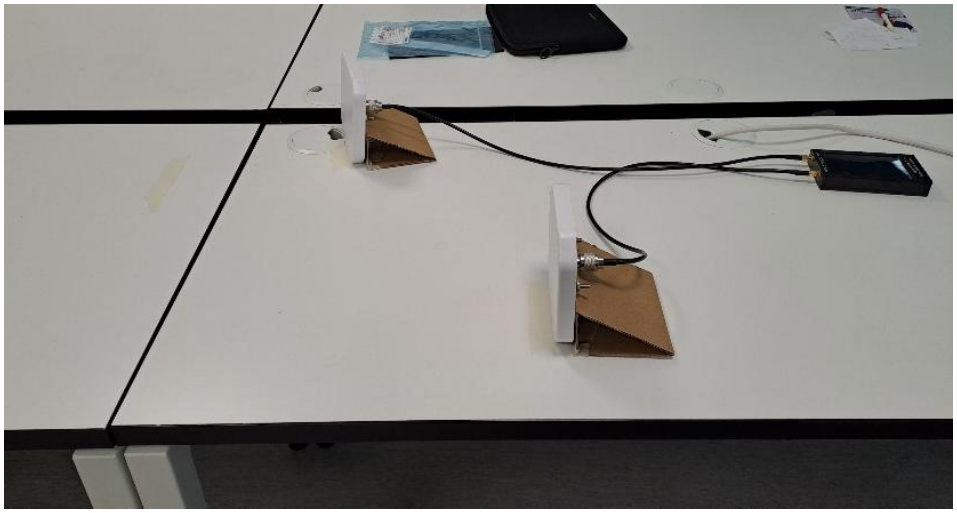

Figure 9. Measurement setup.

As is shown by [22], the antennae radiate in roughly a 160^{o} forward facing angle based on the provided radiation pattern. To avoid the transmission signal bleeding directly into the receiving antenna, the receiver was placed slightly behind the transmitting antenna. The transmitting antenna was then turned to face slightly towards the receiving antenna so that the measured object would be directly in front of the centre point of the transmitting antenna. Finally, this antenna alignment was validated using the S_{21}-parameter value shown in the VNA by making sure it is displaying a large negative gain, which suggests low amounts of received microwave power.

The distance between the measured object and the antennae were maintained using a line of tape on the table such that no part of the object would go over that line during measurement. After fastening the antennae to the table the distance between the receiving antenna and line of tape was measured

to be 47 cm. Because the transmitting antenna was turned slightly the distance between the transmitting antenna and the line of tape was 25 cm from the closer corner and 27 cm from the farther corner.

Despite [5] using variable measurement distances in addition to variable measurement angles in their study, the decision was made to not vary distances in this experiment. While distance has undeniable effects on the power density of the reflected signal, adding distance variation to the collected dataset would have made investigating the effects of RCS on reflections more complex by introducing another source of variation for measured echo power. Additionally, this experiment introduces variable material classes and attempts to create a large diversity in the material sub-types and shapes of the measured objects. Therefore, varying the measurement distance was deemed out of scope for this experiment.

During the data measurement process each object is measured from multiple different angles due the assumption that unique RCS will be occur when altering the object's alignment with regards to the antennae. A different incident angle and a different object profile can cause the signal to be reflected differently, leading to variations in how the received signals behave. The core assumption is that these alterations are significant enough that rotating the object will noticeably alter the power of the received signal. The validity of this assumption is evaluated later in the data exploration subchapter. Objects are rotated such that angles that could be perceived as repeating a previously measured RCS are avoided. In practice this means avoiding very similar object profiles and rotation angle combinations when measuring. The result is that very symmetrical objects have less scans taken of them than objects that are very asymmetrical. Figure 10 has image representations of how the items were rotated.

Figure 10. Examples of object rotation during data measurement.

The measurements were taken in a basement study lounge area at the University of Turku campus. There was no way to limit microwaves from echoing off of the environment but given that the space behind the transmitting antenna was not altered throughout the measurement process, these microwave echoes can be considered a systematic source of error on the experiment. The used space was also large and ventilated meaning that significant water vapour buildup was unlikely to occur during measurement. As was stated previously microwave signals can be attenuated by moisture in the air and if the measurement space was small and badly ventilated. Prolonged human presence in the measurement area could cause early measurements to differ from later measurements due to attenuation. The space was not exclusively used by the measurement personnel throughout the process. Other users of the space were asked to stay clear from the area in front of the antennae and there were no recorded incidents where a measurement was interfered with by another user of the space.

3.3 The Dataset

The data measurements were completed on the 10th of April in 2025 in a single 9-hour session. The resulting dataset contains 1113 samples measured from 34 metal, 33 plastic and 30 cardboard objects. Each object in each class has been scanned a variable number of times, depending on its shape. The total number of metal samples is 373, the number of plastic samples 375 and the number of cardboard samples 365. After an initial round of data exploration it was discovered that one scan from object 1 in the cardboard item samples was a clear measurement error and was therefore removed from the dataset before performing further analysis. Thus, the number of samples in the

analysed dataset is 1112. More information regarding the number of samples taken from objects can be found in table 1. A complete list of the number of scans taken from each object in each class can be found in the appendixes section. Visual demonstration of the outlier value against the other samples measured from object 1 of the cardboard set can be found in figure 11.

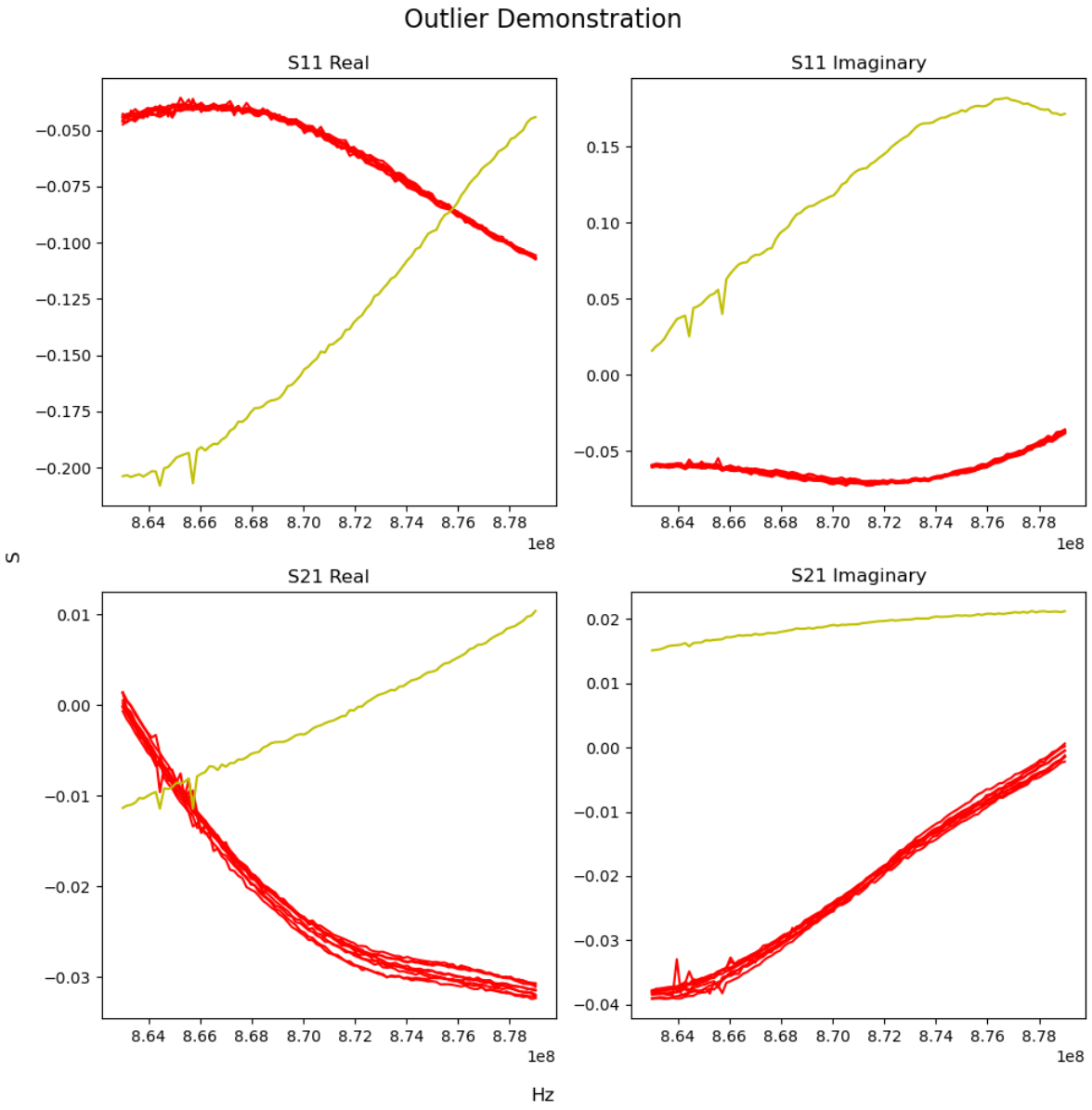

Figure 11. Outlier measurement against other measurements from object 1 of cardboard set. Outlier measurement show in yellow.

Table 1. Class-wise distribution of samples.

Class	Min	Max	Mean	Median	Samples in class
Metal	6	15	10.9706	11.5	373
Plastic	7	16	11.3636	11	375
Cardboard	5	22	12.1333	11.5	364
Overall	5	22	11.4639	11	1112

Each sample is saved in touchstone files which contain the complex valued reflection and transmission coefficients S_{11} and S_{21} over the measurement range. Table 2 shows the first 8 rows of the 1st scan of the 6th metal object as an example. Each file contains 101 rows of data taken in 0.16 MHz intervals.

Table 2. Example rows from touchstone file.

Frequency	S11 Real	S11 Imaginary	S21 Real	S21 Imaginary
863.00 MHz	-0.061454	-0.01752	0.000035	-0.023683
863.16 MHz	-0.060162	-0.016608	-0.000175	-0.023779
863.32 MHz	-0.058163	-0.016598	-0.000543	-0.023736
863.48 MHz	-0.057378	-0.015956	-0.00125	-0.023532
863.64 MHz	-0.056928	-0.015531	-0.001498	-0.023527
863.80 MHz	-0.054795	-0.015381	-0.002001	-0.023404
863.96 MHz	-0.053416	-0.013475	-0.002193	-0.023244
864.12 MHz	-0.049604	-0.019159	-0.001666	-0.026608

3.4 Data Exploration

The visualisations in this chapter were created with Python 3.12.7 in a Jupyter Notebook environment using numpy and matplotlib libraries. The graphs in figure 12 show the S_{21}-parameter values, both real and imaginary, against measurement frequency grouped by object class. The lines in the graph are opaque in order to better visualize how densely packed certain sections of the graph are. It can be very clearly seen that plastic samples form a very tight band of measurements in both the real and imaginary parts of the S_{21}-parameter. The cardboard values can be seen having some outlier measurements that push above and below the plastic band, especially in the imaginary values but seem to mostly fall underneath the plastic band. The clearest distinction is seen in the metal samples, which show a much larger variation in measurement values than the plastic and cardboard samples.

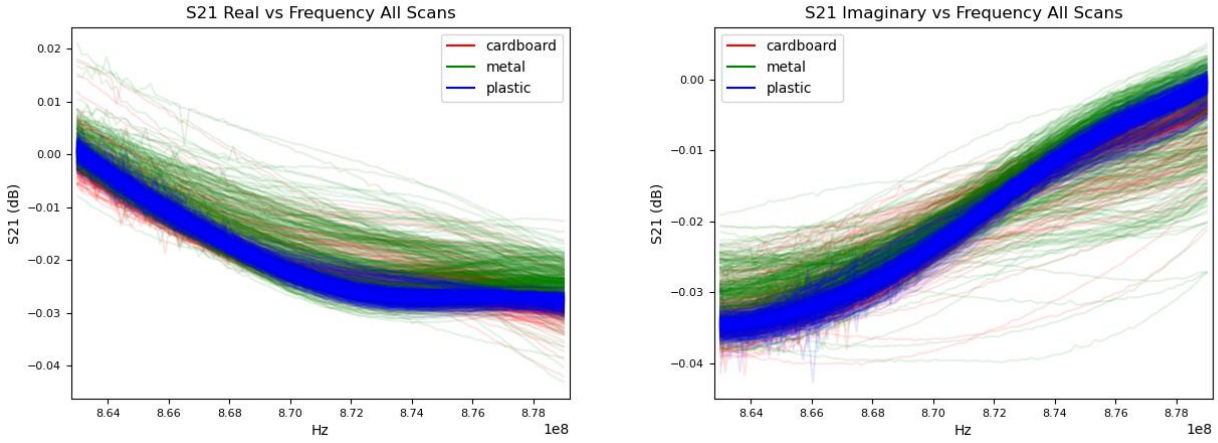

Figure 12. Left: Real component of S21 against frequency. Right: imaginary component of S21 against frequency. Line colours by object class.

As the touchstone files automatically record both the S_{21}- and S_{11}-parameters, the real and imaginary components of the S_{11}-parameter were also plotted against measurement frequency like in figure 12. The expected result was that all measurements would be equal, as S_{11} measures the reflected power from the antenna to the VNA. However, figure 13 clearly shows that for both the real and imaginary parts, metal samples show a very clear variation in measurement values while plastic and cardboard samples show the expected tight band of measurements. Also, as could be seen in the S_{21}-parameter values, cardboard samples have the occasional outliers that push above or below the tight measurement band. Why this variation with metal samples occurred is unclear. The primary hypothesis is that, because measured objects will also scatter radiation back to the transmitting antenna, this backscatter radiation bleeds into the S_{11} measurement, causing the reflected signal strength to increase. This could mean that the transmitting antenna would be enough to measure reflected characteristics by itself in future work.

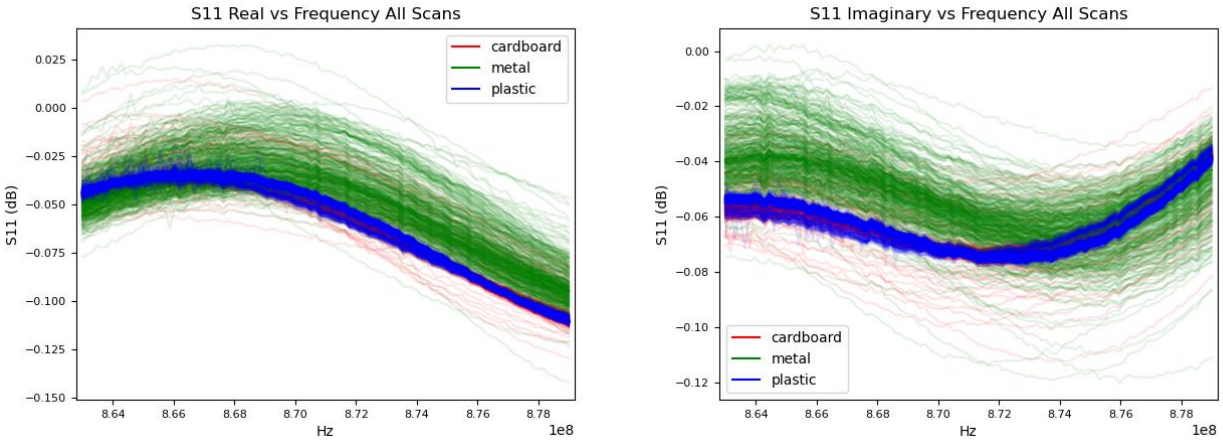

Figure 13. Left: Real component of S11 against frequency. Right: imaginary component of S11 against frequency. Line colours by object class.

The scans from each object were also plotted individually to test the hypothesis that rotating the objects with regards to the transmitting antenna changes the measured values. For this hypothesis to hold one would expect different measurements from the same item to have varying curve shapes or placements. Each graph's y-axis has been normalized to better compare the curves across items and material types.

Throughout this section only the real part of the S_{11} measurements for each object class are shown due to the high visual correlation present across the parameters and parts. Other S-parameter measurements are presented in the appendices section. The previous figures 12 and 13 allude to the result that can also be seen very clearly from these object-wise plotted measurements of plastic samples shown in figure 14. Plastic objects follow a very strict band with almost no variation on the shape or placement of the curves in either the S_{11} or S_{21} values between different objects. This graph also clearly disproves the assumption that rotation would have a noticeable impact on measurements, at least in the context of plastic objects, as there is very little to no variation between measurements from a single object. It can also be seen that for many objects in the set noise can be seen in the first half of the measured values as randomly occurring small spikes and dips. This noise is most clearly visible in the real part of the S_{11} parameter values.

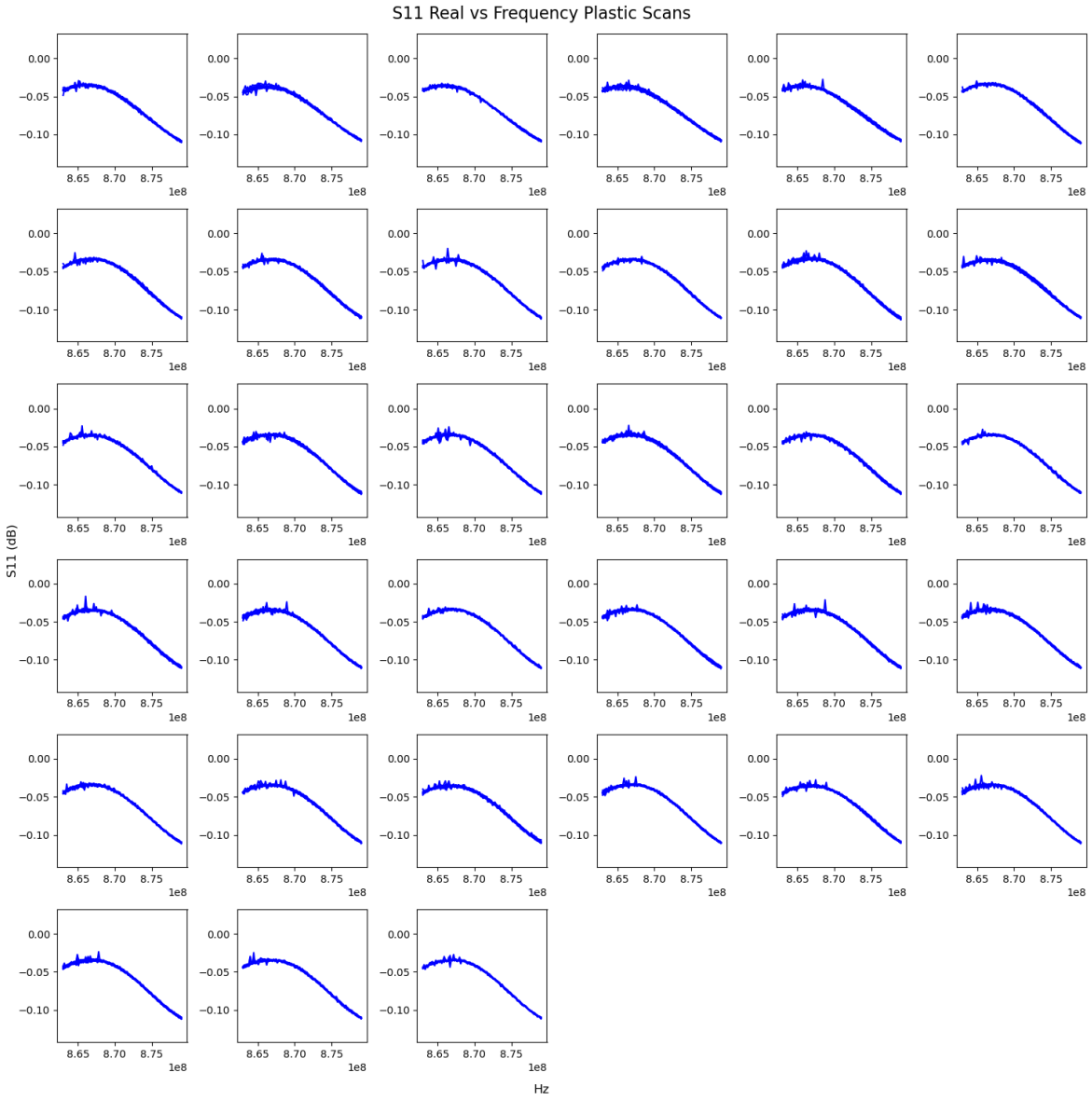

Figure 14. Measurements of the real part of the S11 parameter for plastic samples sorted by item.

Regarding the cardboard objects, most items in the set follow very similar curves to the plastic samples both for S_{11} and S_{21} even down to the noise patterns being best visible in the first half of the real component of S_{11}. Due to the similarity of these near-identical measurement values for plastic and cardboard objects it would be likely that these objects either did not interact with the incident radiation in a meaningful way or have extremely similar absorptive and reflective behaviour with the radiation. Given that there are also items in the cardboard dataset that seem to have varying measurement data based on the rotation of the object (as can be seen from the in-item variation of the measurement curves for items 2, 6, 9 and 10 presented in figure 15) it is unlikely that this similarity is caused by nearly identical interactions.

To more thoroughly explain the previous statement the following chain of logic is presented. The variations in the previously listed items are unlikely to be caused by random chance as, if they were, one would expect items in each of the datasets to have at least some randomly occurring curves that do not fit the visible trends. This randomness cannot be seen in any of the 375 plastic measurements or any of the non-varying cardboard measurements. When varying measurements do occur for an item, most of the measurements for that item show varying curves. Therefore, it can be reasonably assumed that some property of these objects caused incident microwaves to behave differently based on rotation. Thus, some interaction between the object and the incident radiation has to have occurred in these cases.

For the theory to be true that similar measurements are caused by similar reflective and absorptive behaviour, it would be reasonable to assume that rotation would also have an impact on the measured values as it is likely that these in-item variations are caused by item-radiation interaction. For the data to support this interpretation the very similarly curving items in the plastic and cardboard sets would also have to have in-item variation. However, most measurements in the plastic and cardboard sets show almost identical curves and values for each measurement per item. Therefore, it is more likely that these items simply did not interact with the incident radiation.

As a final note, this chain of logic is presented based on observations of the data and not rigorous scientific testing. There can be alterative explanations as to why these phenomena are present and in order to gain definitive proof more tests would have to be performed. However, based on the evidence shown by the data these conclusions provide a reasonable explanation to the observed patterns.

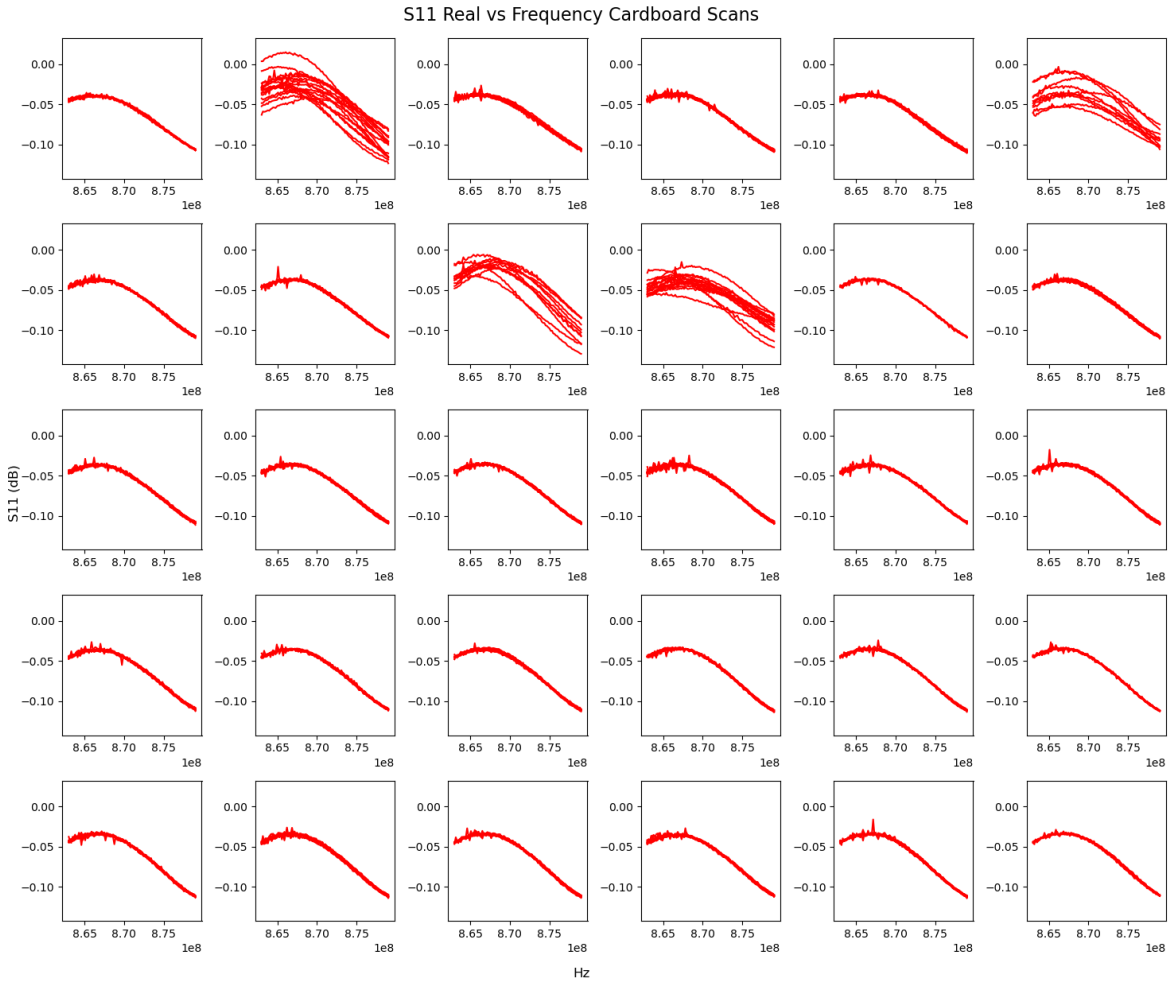

Figure 15. Measurements of the real part of the S11 parameter for cardboard samples sorted by item.

The metal samples seem to further bolster the theory that some items interact with the radiation, and some do not. Items here demonstrate rotational variation with only a few exceptions. But even these exceptions appear to have some small of spread between measurements. No measurement set is situated in as tight a band as non-varying cardboard or plastic samples. Good examples of low rotational variation are items 9, 31 and 34 in the metal set which can be seen in figure 16.

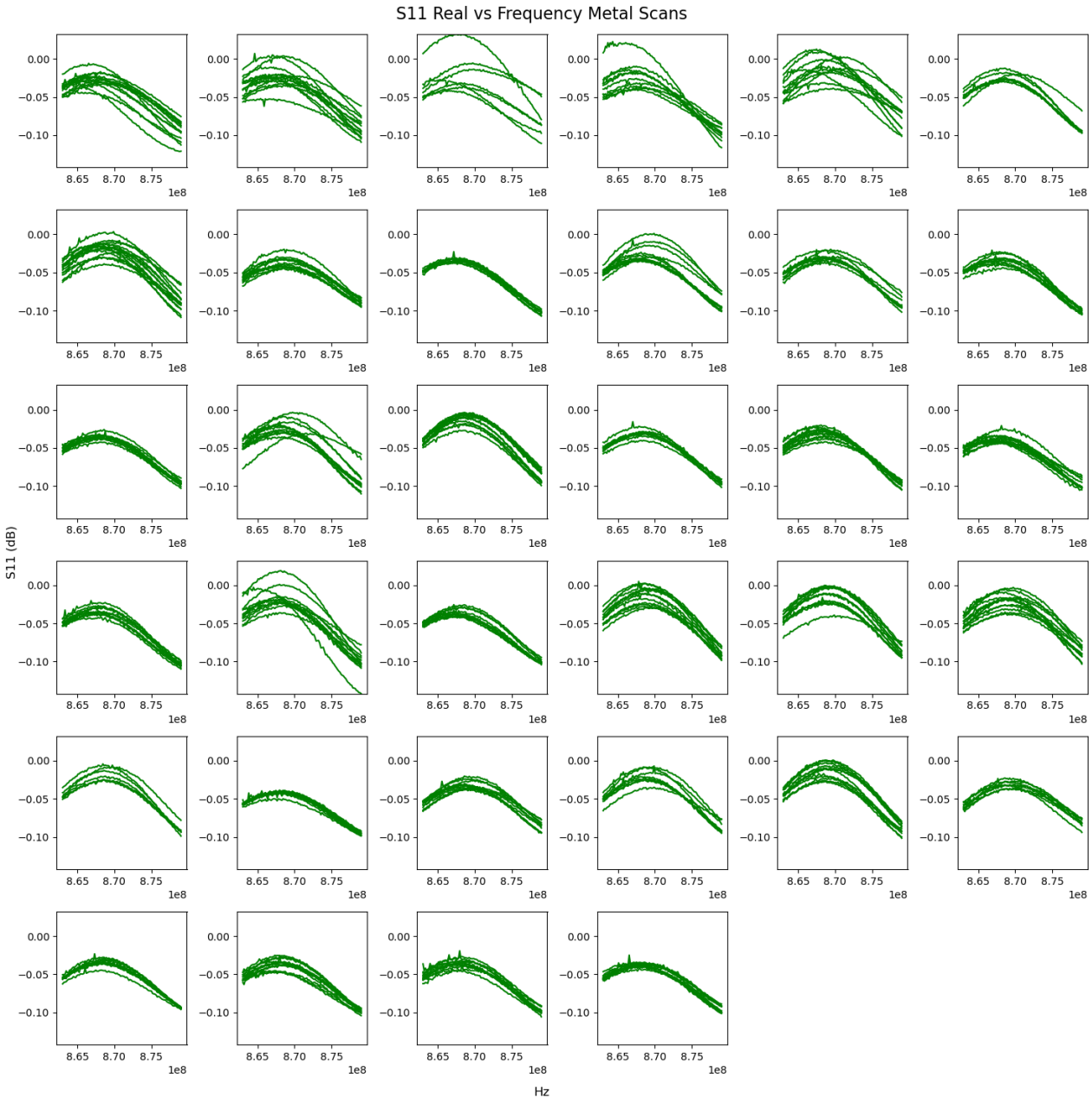

Figure 16. Measurements of the real part of the S11 parameter for metal samples sorted by item.

Due to an oversight in planning there was no recording of which scans correspond to which real world item or which scan was taken from which angle. Therefore, there is no way to verify how exactly rotation or object size or subtype have affected the measurements. This oversight will be corrected in future work.

Based on these visualizations the expected performance of a ML solution would be to proficiently classify metal samples from the other two classes, but there will likely be issues with differentiating plastic from cardboard samples. Similarly the cardboard samples that do show in-item variation

might become difficult to differentiate from metal samples due to their low quantity and similarity with metal samples.

4 Machine Learning Model Training and Evaluation

This chapter examines the process and results of training different DL models on the microwave reflection dataset. Theoretically, these models attempt to classify objects by determining the RCS from the provided EDP measurements.

All model development, including hyperparameter optimization and performance evaluation, was performed in Jupyter Notebooks with Python 3.12.7. The technical stack included Keras for modle creation and sklearn for dataset preparation and performance metric calculation. Confusion matrices were generated using matplotlib and seaborn. Each presented case was managed in an individual notebook.

4.1 Training configurations

In order to test the performance of the models on the dataset thoroughly, four different configurations of the training data are examined:

1. 3-class classification with item-based splitting
2. 3-class classification with random splitting
3. Binary classification with item-based splitting
4. Binary classification with random splitting

In configurations where data is split based on items, the samples are presented to the DL models such that they can have no prior information about items in the validation or test sets. In practice this means that if item x is split into the validation set, all n samples taken from item x are put into the validation set. This simulates an environment in which new objects are meaningfully different from learned ones. By removing complete items from the training set the trained model has to be able to make predictions about items it has no prior information on. Additionally, to make sure that each set will receive an even distribution of each target class, the data splits in these configurations are stratified, which means that the class distributions of the train, test and validation sets will match the distribution of the original dataset.

In the random split cases samples are split without any regard to which object they are from. This simulates the opposite environment where new objects are not meaningfully different from learned objects. By disregarding which object the sample comes from all data samples are effectively

considered to be unique objects. This would mean that even if samples from a single object would be included in both the training and test set information about the test set would not bleed into the training set. Based on the visualizations presented in Chapter 3, all plastic objects display very similar profiles, as do the majority of the cardboard samples. Consequently, it is reasonable to treat each scan of these two classes as equivalent to an individual object. In contrast, the metal class items indicate that there is item-specific variation. However, many individual scans among different items share similarities. This makes it challenging to definitively ascertain which configuration better reflects real-world contexts. Therefore, both configurations will be tested to explore the upper and lower performance boundaries of the microwave classification system.

Based on analysis during exploration, the visualisations of the datasets show that differentiating between cardboard and plastic samples will likely be very difficult. Thus, both multiclass and binary classification are explored. The target variables in the binary classification case will be metal and non-metal as opposed to metal, plastic and cardboard which are present in the multiclass configuration. It is very possible that the system as presented is not a valid way to differentiate plastic from cardboard, but could be competitive in classifying metals from non-metals.

It is good to note that the results from these four configurations aren't directly comparable, as these systems test different contexts. It is very likely that model performance will be best in the binary classification random split context as that configuration is in theory an easier task compared to the rest as it receives a more diverse set of training data and has less classes to predict. Similarly, performance will likely be worse in the item-based split multiclass configuration as that is theoretically the hardest task for the opposite reasons.

In each case the training process is such that first two sets of hyperparameter optimization are performed. The goal of the first round of hyperparameter optimization is to find which of the four general hidden layer configurations and optimizers is best. These configurations attempt to provide a diverse set of general structures, although the exact number of neurons in each layer and number of layers are arbitrary. The four tested hidden layer structures are presented in table 3.

Table 3 General layer structures.

Small number of layers, small layer size	Small number of layers, large layer size	Large number of layers, large layer size	Large number of layers, large layer size
1 x 128 neurons	2 x 512 neurons	2 x 128 neurons	6 x 512 neurons
2 x 64 neurons	1 x 256 neurons	6 x 64 neurons	2 x 256 neurons

The tested optimizers are Adam and RMSprop. SGD was not utilized due to its reliance on case-specific learning rate, decay rate and momentum tuning [29]. That level of fine-tuning was deemed unnecessarily complex for this thesis. Each combination is trained for 400 epochs and the best performing model-optimizer combination is chosen based on validation set accuracy for multiclass cases and validation set loss on the binary cases. The choice of optimizing for validation loss in binary cases was done due to the binary random split configuration validation accuracy quickly reaching 100% during training. The loss function used in the multiclass cases is categorical cross-entropy, a variant of cross-entropy usable in multiclass classification and the loss function used in the binary cases is binary cross-entropy, which was presented in chapter 2.1.2. The best performing model is tested against a test set to obtain a test set accuracy rating for the purposes of model selection.

The second round of hyperparameter optimization will use the best performing layer structure and optimizer and fine tunes the number of layers to find the optimally performing model based on validation accuracy in the multiclass cases and validation loss in the binary cases. More detailed descriptions of layer count variations can be found from table 4. In this round each model is trained for 500 epochs instead. The fine-tuned model is also tested on the same test set to obtain an accuracy rating for model selection purposes.

Table 4. Round 2 hyperparameter optimization layer count tuning.

Small number of layers, small layer size	Small number of layers, large layer size	Large number of layers, large layer size	Large number of layers, large layer size
0-1 x 256 neurons	0-1 x 1024 neurons	0-1 x 256 neurons	0-1 x 1024 neurons
1-2 x 128 neurons	1-3 x 512 neurons	1-3 x 128 neurons	3-6 x 512 neurons
1-3 x 64 neurons	1-2 x 256 neurons	3-6 x 64 neurons	1-3 x 256 neurons

The model with the best test set accuracy is then used for performance estimation. The performance estimation step is done by running the best performing model configuration through K-fold cross-validation with k=10. Therefore, one fold is used as the test set for evaluating the performance for that fold and the remaining nine are being split into training and validation data in a 90%/10% ratio. The validation set is used to optimize the number of epochs the model is trained for. Performance metrics are collected and averaged across the folds. The performance metrics used in this experiment are accuracy, precision, recall and F1-score. A confusion matrix is presented for each optimization step and for the performance estimation step. A summary of the best performing

models from each set of hyperparameter optimizations performed in the following subchapters is shown in table 5.

Table 5. Summary of hyperparameter optimization results.

Case	General layer structure	Optimizer	Detailed Layer structure	Test set Accuracy
1	Large network, small layers	RMSprop	2x128 + 6x64	94.07%
2	Small network, small layers	Adam	1x128 + 2x64	94.64%
3	Large network, small layers	RMSprop	3x128 + 4x64	87.29%
4	Large network, large layers	Adam	4x512 + 3x256	100.00%

4.2 Model Selection and Performance Estimation

4.2.1 Case 1: Item-based Multiclass

In the first hyperparameter optimization round the RMSprop -optimizer consistently outperformed the Adam-optimizer. However, the differences were not very large in terms of absolute performance. Table 6 shows the full hyperparameter optimization results for round 1. The best performing structure and optimizer combination for this round was the large network, small layers with RMSprop -optimizer at 98.87% accuracy. The best performing model attained a test set performance of 94.07% with the highest level of misclassification interestingly occurring between metal and cardboard samples rather than cardboard and plastic samples as shown in figure 17.

Table 6. Case 1 round 1 hyperparameter optimization results.

Layers	Optimizer	Validation Accuracy
Large network, small layers	RMSprop	98.87%
Large network, large layers	RMSprop	97.75%
Small network, large layers	RMSprop	97.75%
Small network, small layers	RMSprop	97.75%
Large network, small layers	Adam	96.63%
Small network, large layers	Adam	96.63%
Small network, small layers	Adam	96.63%
Large network, large layers	Adam	96.63%

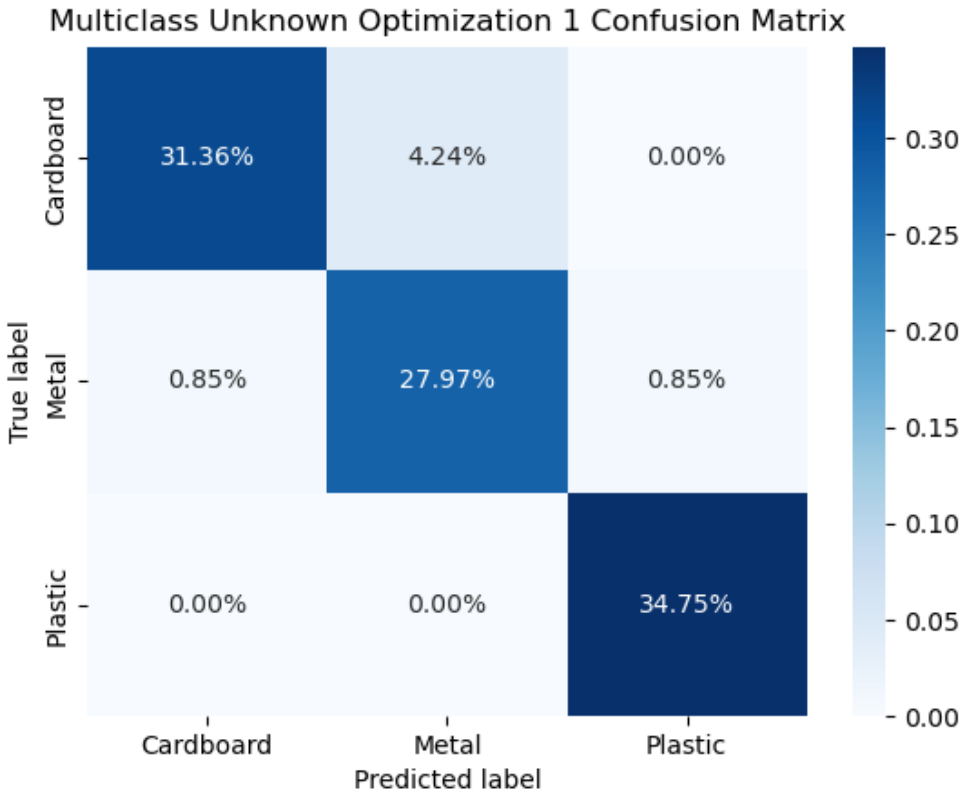

Figure 17. Case 1 hyperparameter optimization 1 confusion matrix for best performing model. Unknown in the title refers to item-based splitting.

The second round of hyperparameter optimization shows that the model performance seems mostly agnostic to finer layer tuning with all but 1 of the top 10 models performing equally well. The best performing model managed to attain a 100% accuracy on the validation set but in fact underperforms on the test set compared to the first round model with a test set accuracy of 88.98%. Misclassifications occurred much more frequently between metal and cardboard and some cases even with cardboard and plastic as well. Test set classification results are shown more precisely in figure 18 and full hyperparameter optimization results for round 2 in table 7.

Table 7. Case 1 round 2 hyperparameter optimization results.

Large layers (256 neurons)	Medium layers (128 neurons)	Small layers (64 neurons)	Validation Accuracy
1	2	6	100%
0	2	5	98.87%
0	1	4	98.87%
1	2	3	98.87%
1	3	6	98.87%
0	2	6	98.87%
1	3	4	98.87%

Large layers (256 neurons)	Medium layers (128 neurons)	Small layers (64 neurons)	Validation Accuracy
1	2	4	98.87%
0	1	6	98.87%
1	2	6	98.87%

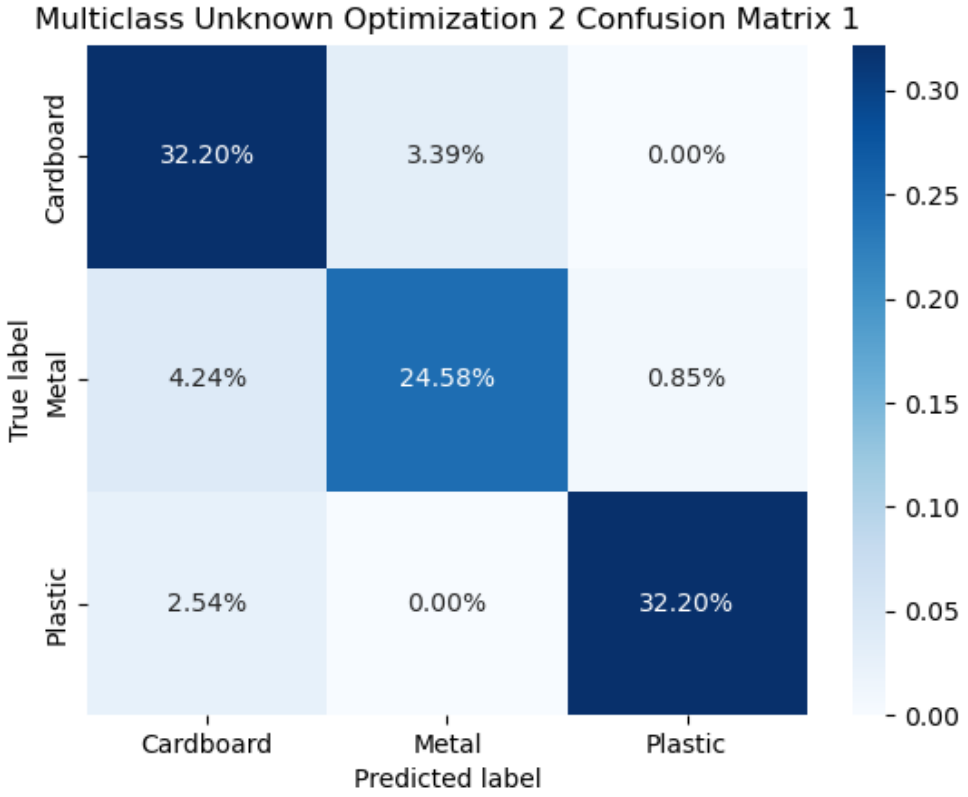

Figure 18. Case 1 hyperparameter optimization 2 confusion matrix for best performing model. Unknown in the title refers to item-based splitting.

Due to the best performing model underperforming the first optimization round performance, the second best model from this round was also evaluated on the test set. It received an accuracy rating of 77.12% mainly due to its to atrocious performance on the cardboard samples which can be seen in figure 19. As both models tested from the second round of hyperparameter optimization had worse performance on the test set than the best first round model, the model chosen for K-fold cross-validation was the best performing first round model.

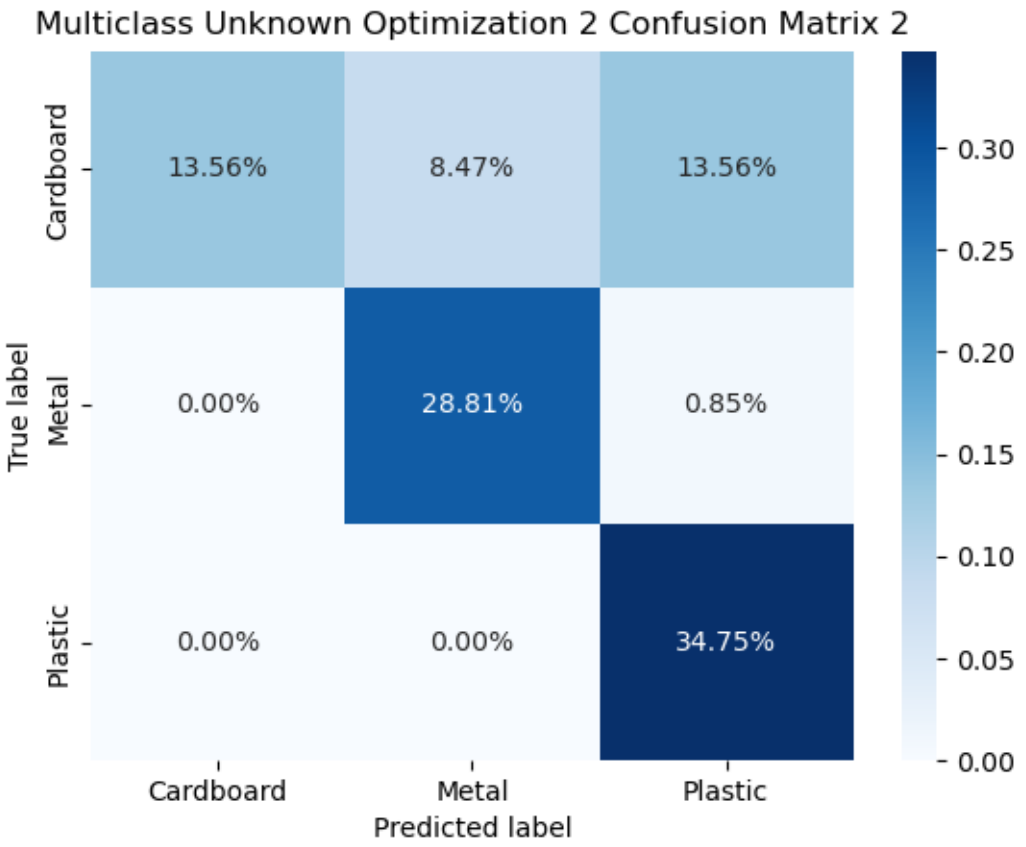

Figure 19. Case 1 hyperparameter optimization 2 confusion matrix for model with second best performance. Unknown in the title refers to item-based splitting.

The 10-fold cross-validation results for case 1 are reported in table 8. Overall the model performed surprisingly well with an average accuracy of 83.31% and a solid F1-score of 0.8333. There exists a difference between accuracy reported by the confusion matrix and average accuracy of 0.05% caused by rounding. While the hyperparameter optimization stage confusion matrices showed that the likeliest misclassifications were to occur between metal and cardboard samples, the final aggregate confusion matrix in figure 20 clearly shows that the most likely misclassifications happen between plastic and cardboard samples. To highlight how drastic these differences are, the recall for the metal samples was 94.63%, 71.16% for the cardboard samples and 83.99% for the plastic samples. Additionally, the performance metrics across folds vary significantly with the best metrics row having all values in the 99% and the worst being in the 50-53% range. This variance likely occurs due to the low number of unique items present in the dataset, combined with the few cardboard samples that show rotational variation similar to that of a metal sample causing confusion during training and while present in the test set. As these factors combine, some test sets will simply be much harder than others to predict correctly.

Table 8. Case 1 10-fold cross-validation results.

Fold	Precision	Recall	F1-score	Accuracy
1	86.65%	85.46%	0.8534	86.73%
2	95.00%	93.75%	0.9385	94.64%
3	93.33%	90.09%	0.9048	91.27%
4	55.83%	52.08%	0.5339	52.68%
5	77.48%	74.27%	0.7541	73.95%
6	93.83%	87.18%	0.8866	90.29%
7	99.10%	99.36%	0.9922	99.26%
8	95.77%	94.53%	0.9492	95.00%
9	98.96%	98.92%	0.9892	98.85%
10	53.21%	53.21%	0.5315	50.48%
Average	84.91%	82.89%	0.8333	83.31%

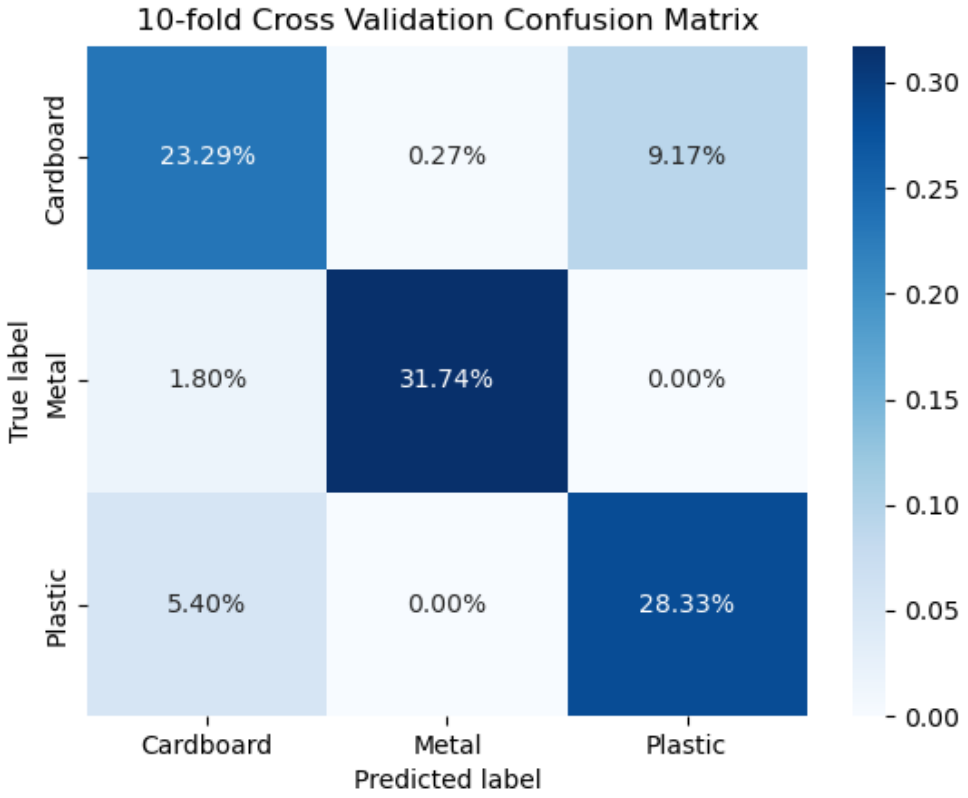

Figure 20. 10-fold cross-validation aggregate confusion matrix for case 1.

4.2.2 Case 2: Random Multiclass

In the random multiclass case there isn't an optimizer that clearly outperforms the other but the performance differences between the best and worst performing models are much larger than in

case 1. Additionally, the margin between the best and second best is a lot larger than in case 1. The best performing combination is a small network with small layers and the Adam-optimizer. Interestingly this combination performed even better on the test set, receiving and accuracy rating of 95.54%. Full results of round 1 hyperparameter optimization are shown in table 9 and test set confusion matrix for best performing model in figure 21.

Table 9. Case 2 round 1 hyperparameter optimization results.

Layers	Optimizer	Validation Accuracy
Small network, small layers	Adam	94.00%
Large network, large layers	RMSprop	91.00%
Small network, large layers	Adam	90.00%
Small network, small layers	RMSprop	90.00%
Small network, large layers	RMSprop	89.00%
Large network, small layers	Adam	88.00%
Large network, large layers	Adam	88.00%
Large network, small layers	RMSprop	88.00%

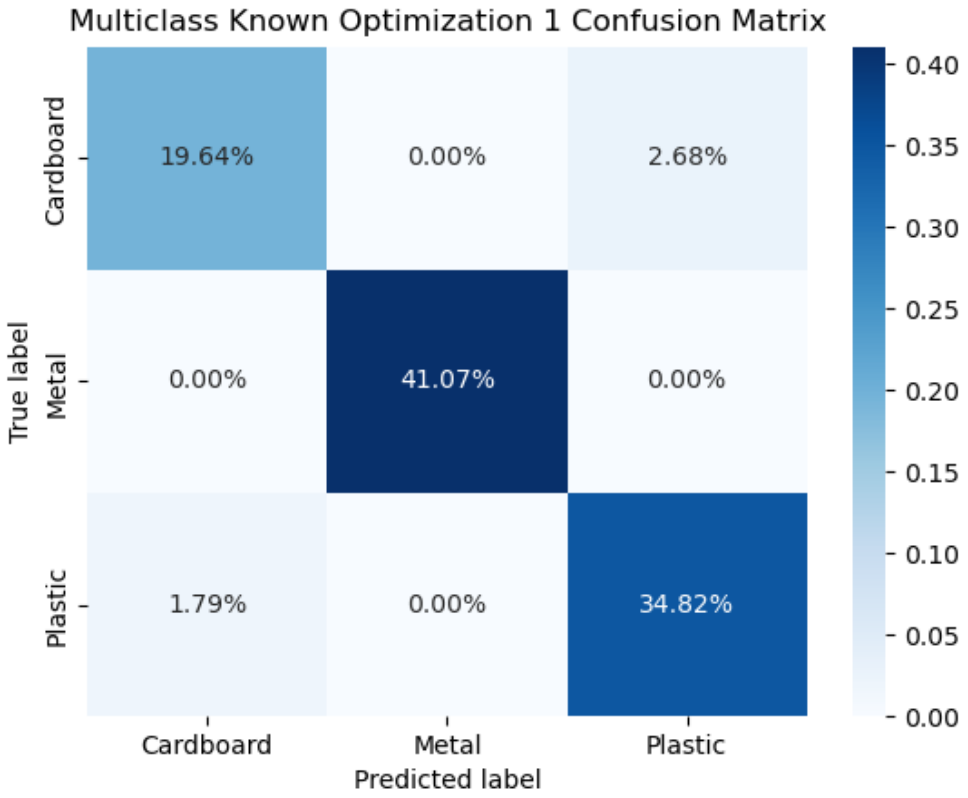

Figure 21. Case 2 hyperparameter optimization 1 confusion matrix for best performing model. Known in the title refers to random splitting.

For the second round of optimization the same small layer configuration of one 128 neuron layer and two 64 neuron layers retains the top spot but loses a percentage of validation accuracy. The same happens with the test set with accuracy dropping to 94.64%. This difference in performance is likely caused by the random initialization of model weights when training begins. Due to the same model structure being the best performing in both hyperparameter optimization rounds it is chosen as the model for K-fold cross-validation. Full results of round 2 hyperparameter optimization and the corresponding best performing model confusion matrix are shown in table 10 and figure 22 respectively.

Table 10. Case 2 round 2 hyperparameter optimization results.

Large layers (256 neurons)	Medium layers (128 neurons)	Small layers (64 neurons)	Validation Accuracy
0	1	2	93.00%
0	2	1	92.00%
1	1	1	92.00%
1	2	2	92.00%
1	1	3	91.00%
0	1	1	91.00%
1	1	2	91.00%
0	2	2	91.00%
1	2	3	90.00%
1	2	1	90.00%

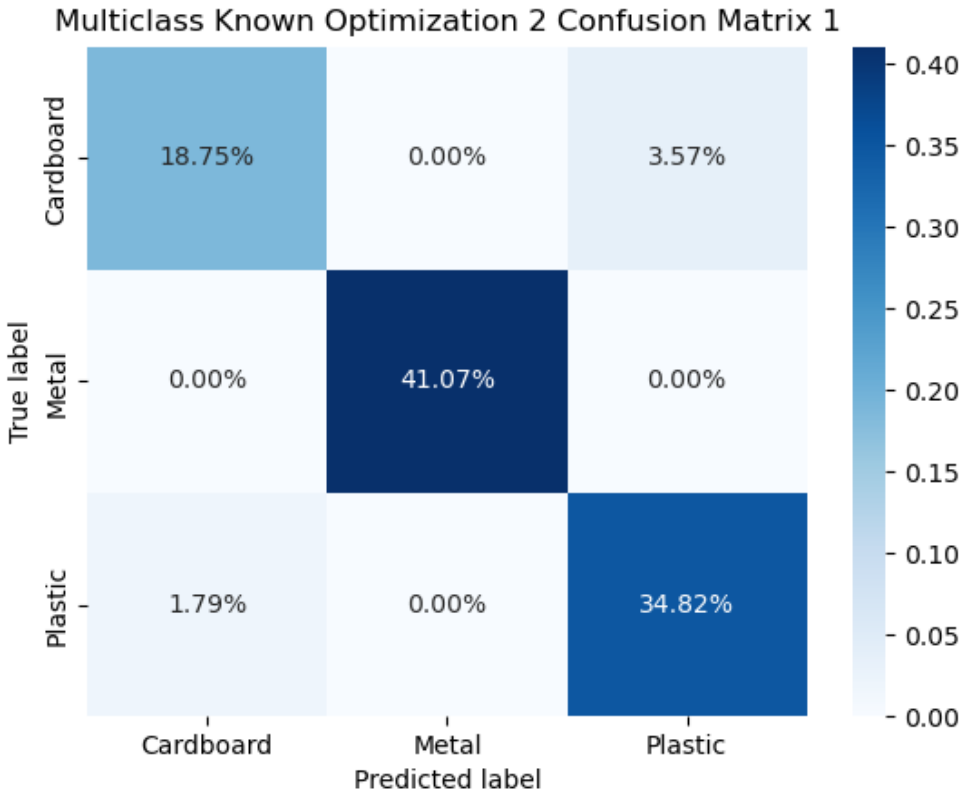

Figure 22. Case 2 hyperparameter optimization 2 confusion matrix for best performing model. Known in the title refers to random splitting.

As is to be expected, the random case average performance metrics are all better than in case 1 and the per fold statistics are much more stable throughout as is shown in table 11. In the confusion matrix shown in figure 23 it can be seen that the vast majority of misclassifications occur across plastic and cardboard and specifically when cardboard is misclassified as plastic. Otherwise the model seems to perform near perfectly with very minor quantities of FP or FN predictions occurring in other sections of the confusion matrix.

Table 11. Case 2 10-fold cross-validation results.

Fold	Precision	Recall	F1-score	Accuracy
1	88.96%	88.17%	0.8698	86.61%
2	88.36%	84.52%	0.8489	85.71%
3	87.45%	85.92%	0.8593	86.49%
4	92.06%	88.17%	0.8894	89.19%
5	85.83%	85.44%	0.8442	84.68%
6	89.54%	88.16%	0.8688	87.39%
7	88.16%	84.87%	0.8477	86.49%
8	90.28%	87.39%	0.8653	87.39%
9	90.43%	88.57%	0.8811	88.29%

Fold	Precision	Recall	F1-score	Accuracy
10	89.59%	85.19%	0.8549	87.39%
Average	89.07%	86.66%	0.8629	86.96%

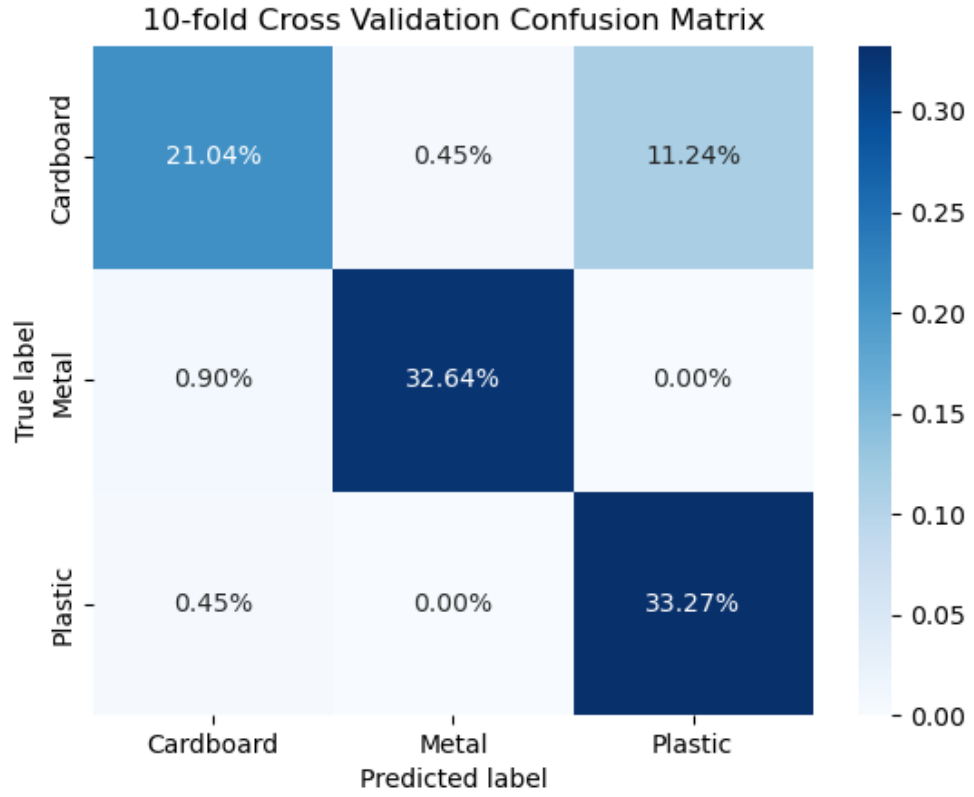

Figure 23. 10-fold cross-validation aggregate confusion matrix for case 2.

4.2.3 Case 3: Item-based Binary

The binary item-based case shows an interesting situation where the best performing model/optimizer combination uses RMSprop but the rest of combinations in the top 4 use only Adam. The model and optimizer combination chosen for round 2 optimization is a large network with small layers and the RMSprop optimizer which received an 86.44% accuracy score with the test set. Given that binary contexts are supposed to be easier to classify, at least in theory, for the models, it is surprising how much worse the test set performance is. One possible explanation for this phenomenon is that the way data was split in this case simply made a much harder test set and the training set was incapable of providing enough variability for predicting harder samples. Full results of round 1 hyperparameter optimization are shown in table 12 and test set performance is shown in figure 24.

Table 12. Case 3 round 1 hyperparameter optimization results.

Layers	Optimizer	Validation Loss
Large network, small layers	RMSprop	5.36e-9
Large network, large layers	Adam	1.28e-8
Small network, small layers	Adam	3.42e-7
Large network, small layers	Adam	1.25e-6
Small network, small layers	RMSprop	1.29e-6
Large network, large layers	RMSprop	1.80-6
Small network, large layers	RMSprop	5.26e-6
Small network, large layers	Adam	1.80e-5

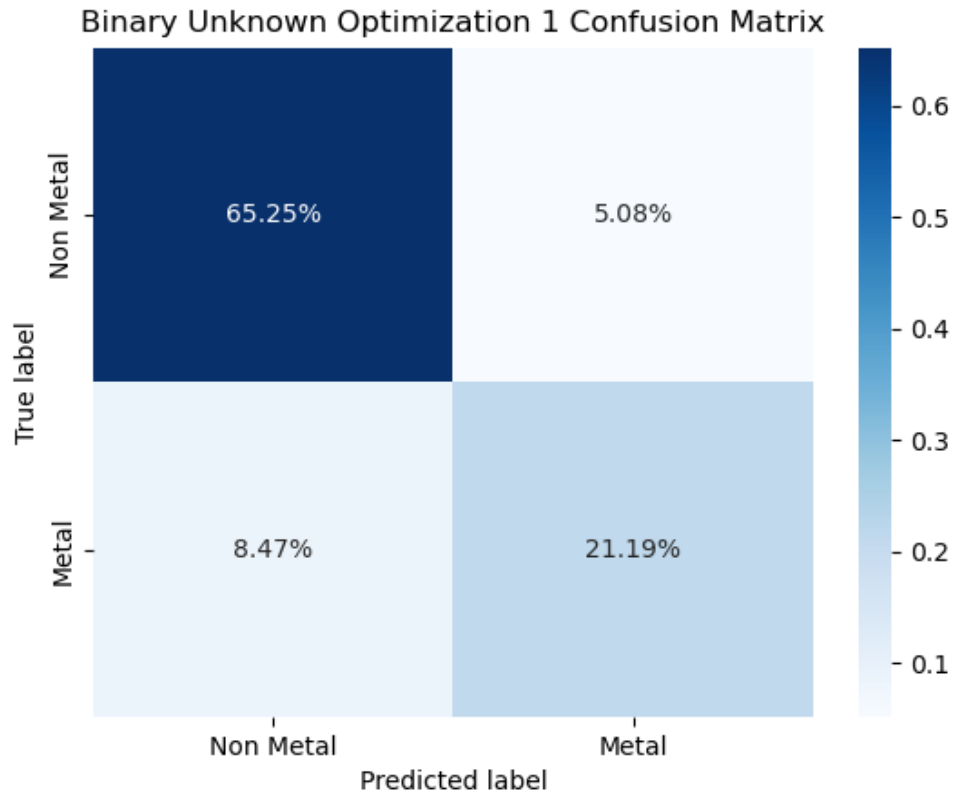

Figure 24. Case 3 hyperparameter optimization 1 confusion matrix for best performing model. Unknown in the title refers to item-based splitting.

In the second round the best performing layer structure changes slightly with one more medium layer and two fewer small layers. While its validation loss is slightly worse than the best performing model from round 1, it attains a better test set accuracy of 87.29% and is therefore chosen as the model for k-fold validation. Full results shown in table 13 and test set performance displayed in figure 25.

Table 13. Case 3 round 2 hyperparameter optimization results.

Large layers (256 neurons)	Medium layers (128 neurons)	Small layers (64 neurons)	Validation Loss
0	3	4	6.81e-9
1	3	3	1.00e-8
1	1	6	1.35e-8
0	2	5	1.44e-8
0	1	6	1.83e-8
1	2	3	2.00e-8
1	2	5	2.12e-8
0	1	5	3.54e-8
1	2	6	4.57e-8
0	2	6	5.49e-8

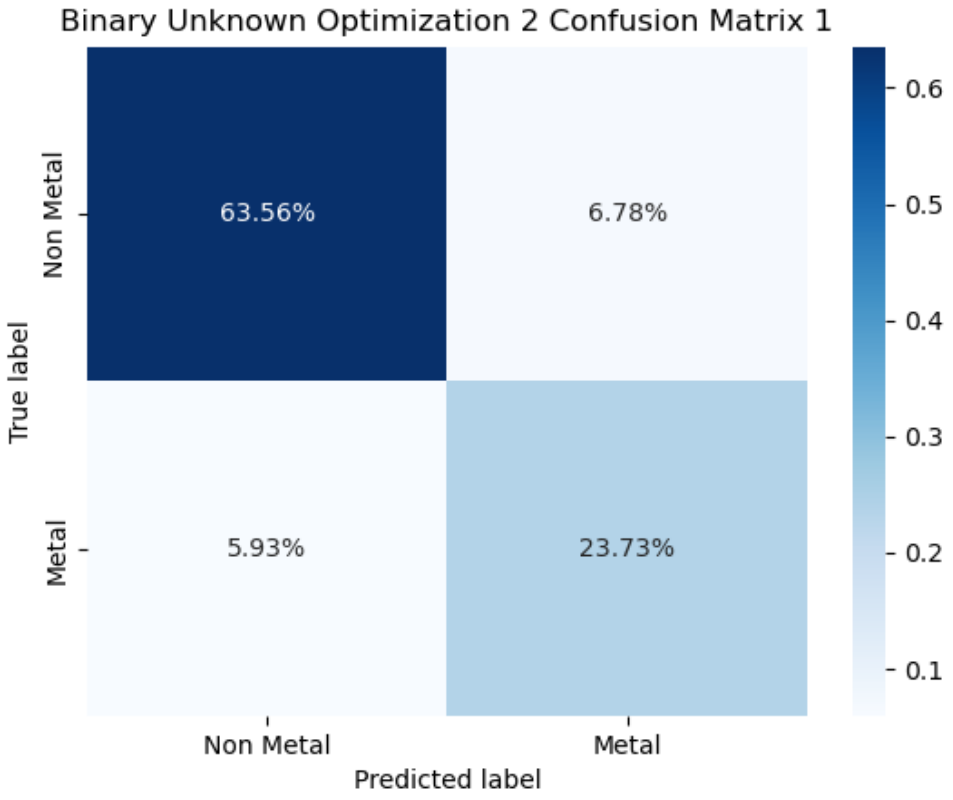

Figure 25. Case 3 hyperparameter optimization 2 confusion matrix for best performing model. Unknown in the title refers to item-based splitting.

Based on the average performance results the binary model has a much better overall statistics than the multiclass models. However, similar performance fluctuations to case 1 can be seen with the best performing folds reaching a perfect 100% in all statistics and the worst being in the 70-76% range. This is likely caused by the same problems as in case 1, those being low number of unique

items and a few confusing cardboard samples. Interestingly the precision and recall across the metal class are worse in this case than case 2. Additionally recall for the metal class is only slightly better and precision much worse than in case 1. This suggests that while the multiclass model is learning to differentiate between the cardboard and plastic samples it also has to put in more effort differentiating the metal samples from the non-metal samples, resulting in better performance in the metal class. Full results shown in table 14 and figure 26. A difference of 0.42% can be observed between the accuracy reported in table 14 and figure 26 due to rounding.

Table 14. Case 3 10-fold cross-validation results.

Fold	Precision	Recall	F1-score	Accuracy
1	97.26%	95.45%	0.9621	96.46%
2	100%	100%	1.00	100%
3	100%	100%	1.00	100%
4	93.96%	92.36%	0.9307	93.75%
5	95.70%	88.24%	0.9109	93.28%
6	100%	100%	1.00	100%
7	92.86%	96.97%	0.9459	95.56%
8	74.95%	76.84%	0.6989	70.00%
9	86.90%	90.18%	0.8702	87.36%
10	100%	100%	1.00	100%
Average	94.16%	94.00%	0.9319	93.64%

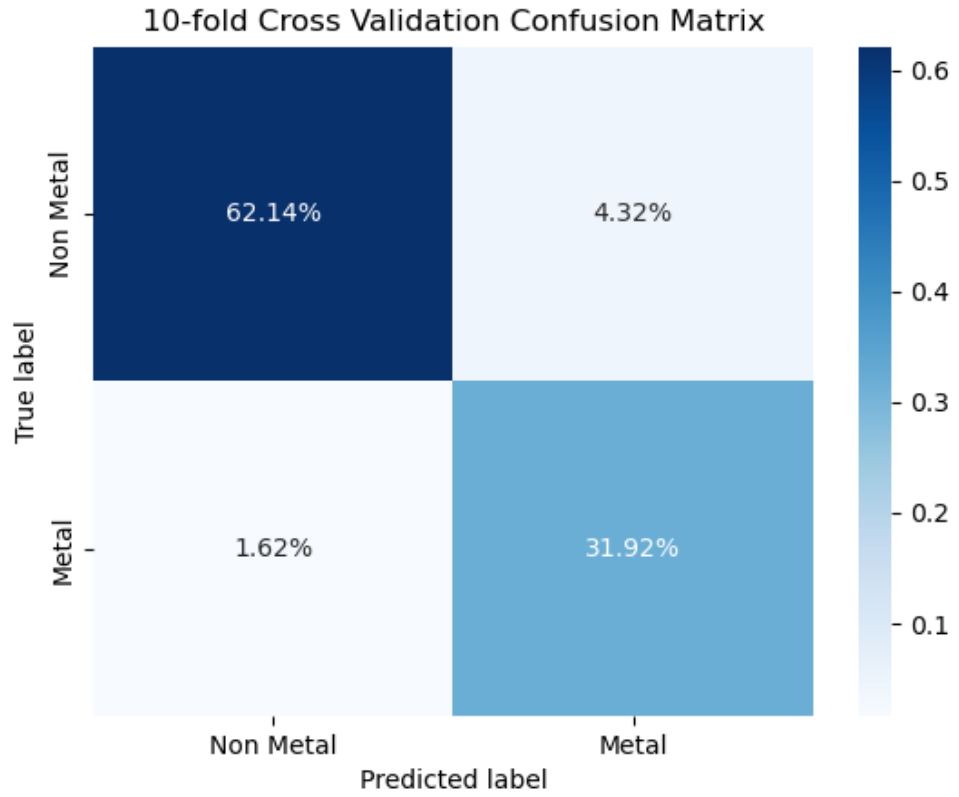

Figure 26. 10-fold cross-validation aggregate confusion matrix for case 3.

4.2.4 Case 4: Random Binary

The random binary case has larger validation losses by orders of magnitude compared to the item-based binary case. Although, it has been shown that the item-based case is much more volatile in terms of performance during training, so these more modest loss values aren't unusual. The best performing model and optimizer combination is a large network with large layers and the Adam optimizer. Its test set performance was an excellent 98.21%. Full round 1 hyperparameter optimization results are shown in table 15 and figure 27.

Table 15. Case 4 round 1 hyperparameter optimization results.

Layers	Optimizer	Validation Loss
Large network, large layers	Adam	0.0501
Large network, small layers	RMSprop	0.0638
Small network, small layers	Adam	0.0802
Small network, large layers	RMSprop	0.0834
Large network, large layers	RMSprop	0.0846
Small network, small layers	RMSprop	0.0847
Small network, large layers	Adam	0.0873

Layers	Optimizer	Validation Loss
Large network, small layers	Adam	0.0879

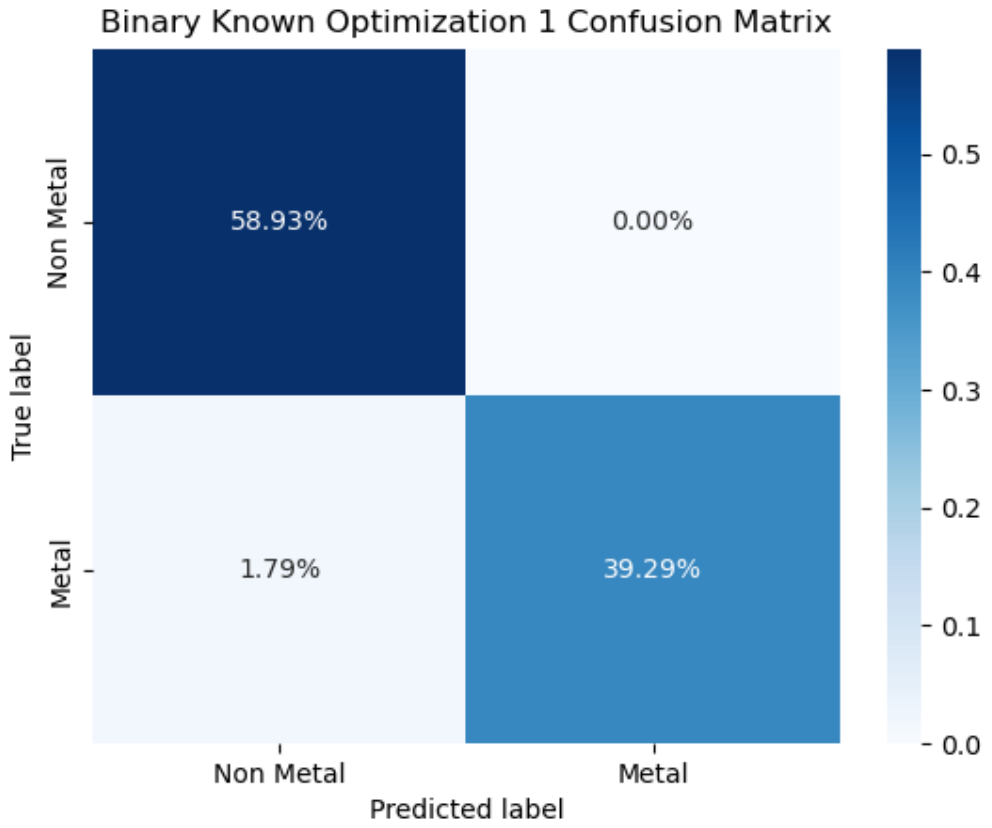

Figure 27. Case 4 hyperparameter optimization 1 confusion matrix for best performing model. Known in the title refers to random splitting.

The best performing model in the second-round underperforms slightly in terms of validation loss compared to the best performing model from the first round but receives a perfect 100% test set accuracy and is therefore chosen as the K-fold cross-validation model. Full results shown in table 16 and figure 28.

Table 16. Case 4 round 2 hyperparameter optimization results.

Large layers (1024 neurons)	Medium layers (512 neurons)	Small layers (256 neurons)	Validation Loss
0	4	3	0.0530
1	6	3	0.0648
0	5	1	0.0698
0	5	2	0.0698
0	6	2	0.0698

Large layers (1024 neurons)	Medium layers (512 neurons)	Small layers (256 neurons)	Validation Loss
0	6	1	0.699
1	5	3	0.0715
0	6	3	0.0730
1	3	3	0.0734
1	6	1	0.0735

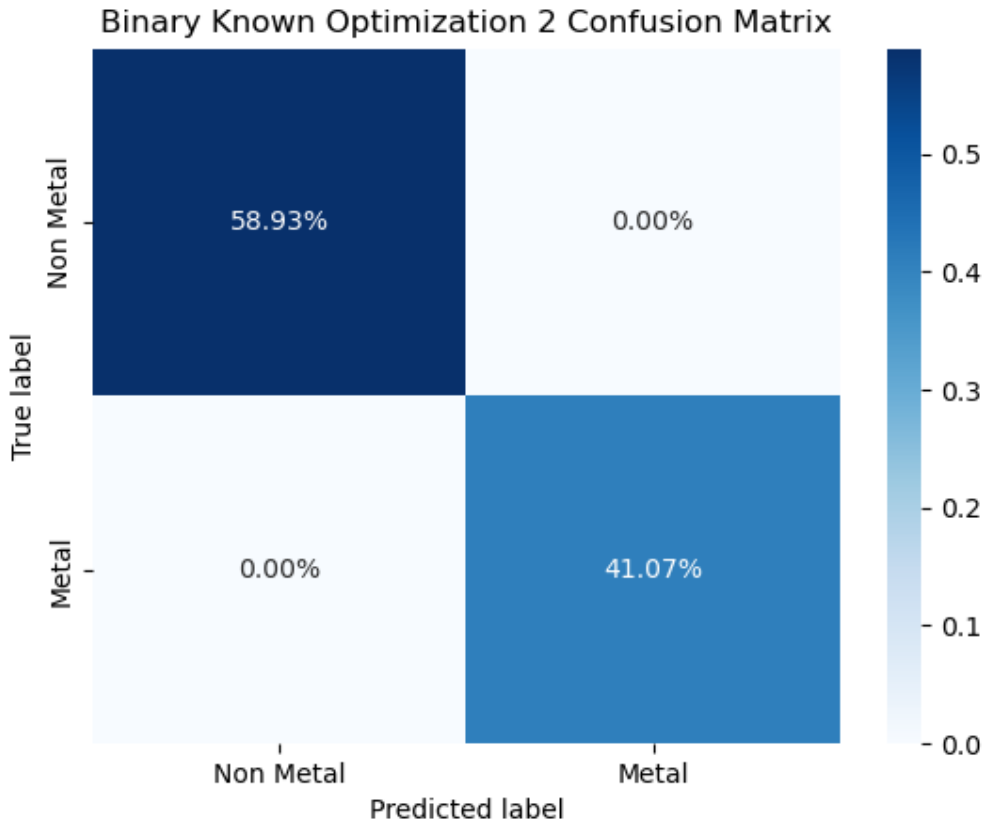

Figure 28. Case 4 hyperparameter optimization 2 confusion matrix for best performing model. Known in the title refers to random splitting.

The 10-fold cross-validation metrics show the same pattern when compared to case 3 that could be seen between case 1 and 2; the volatility of metric results is much lower than in the item-based case across folds. Differences are in the single digits rather than over 20% and the overall performance is a few percentage points higher than in case 3. Recall across the metal class improves massively and is the best across all the cases, but precision still lags the multiclass models. Full results are shown in table 17 and figure 29.

Table 17 Case 4 10-fold cross-validation results.

Fold	Precision	Recall	F1-score	Accuracy
1	94.87%	97.40%	0.9596	96.43%
2	92.31%	96.20%	0.9386	94.64%
3	96.68%	97.30%	0.9698	97.30%
4	94.44%	97.47%	0.9576	96.40%
5	95.24%	97.26%	0.9609	96.40%
6	91.49%	94.44%	0.9241	92.79%
7	94.68%	96.38%	0.9531	95.50%
8	92.55%	95.07%	0.9338	93.69%
9	96.68%	97.30%	0.9698	97.30%
10	94.44%	96.48%	0.9523	95.50%
Average	94.34%	96.53%	0.9520	95.59%

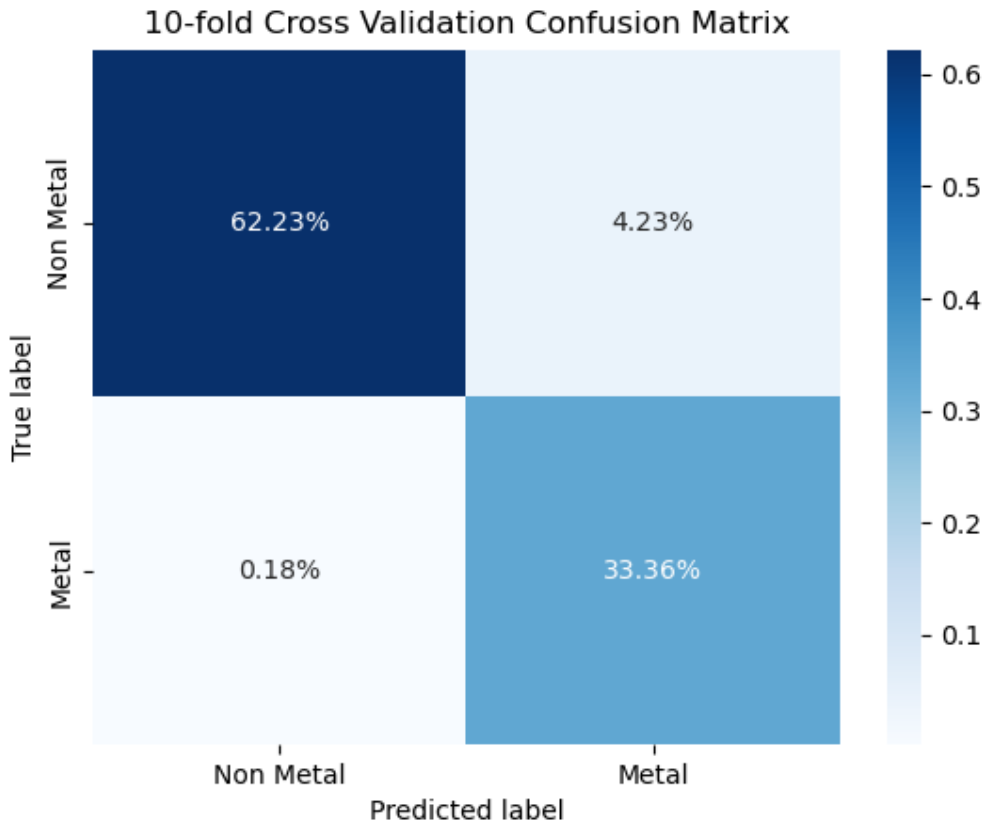

Figure 29. 10-fold cross-validation aggregate confusion matrix for case 4.

5 Discussion

While the different cases are not strictly comparable with each other due to the inherent difficulty differences present in each case, some insights can still be drawn from how the performance metrics change across the different case types. This chapter will take a more detailed look into the performance metric values that were calculated based on the model performances in each case in chapter 4 and discuss their possible implications. All of the class-wise metrics in this chapter are calculated based on the 10-fold cross-validation confusion matrix data found in the previous chapter.

As was predicted before training began, the accuracy of the models increases incrementally when moving from case 1 to case 4 as shown in table 18. This result supports the hypothesis that item-based cases are inherently more difficult than random cases and that binary cases are easier than multiclass cases. What was not obvious was which of the adjustments would have a more significant impact on performance: moving from multiclass to binary or moving from item-based to random. Based on the data this seems to clearly be moving from multiclass over to binary, which gives a 10.33 percentage point increase in accuracy in the item-based set and an 8.63 percentage point increase in the random set with similar increases being found in other performance metrics as well. Comparing these to the performance gains from the item-based to random transition, which are 3.65 percentage points in the multiclass case and 1.95 percentage points in the binary case, shows that the increases in absolute percentage point terms are 3 to 5 times higher depending on the point of comparison when moving from multiclass to binary.

Table 18. Average cross-validation metrics by case.

Case	Precision	Recall	F1-score	Accuracy
1	84.91%	82.89%	0.8333	83.31%
2	89.07%	86.66%	0.8629	86.96%
3	94.16%	94.00%	0.9319	93.64%
4	94.34%	96.53%	0.9520	95.59%

As for why these increases occur, it was clear from the data visualization that the most difficult task for the model will be to differentiate between plastic and cardboard samples. When moving from the multiclass case to the binary case, the requirement for the model to differentiate these classes is removed which will increase the model's performance with the now combined non-metal class. This interpretation is further supported by data in tables 19-21. The reason for the performance

improvement in the random case is less obvious. One possibility is that the assumption of samples being considerable as unique objects is incorrect and information about the test set bleeds into the training set which causes the predicting model to overperform. Another is that when disregarding item-based splitting, the model training receives a more diverse spread of data, allowing for better predictions. At this point it is impossible to state with confidence if either of these two possibilities are true. However, the degree of possible overperformance isn't that significant overall given that the accuracy increases when moving from item-based to random cases are only 3.65 and 1.95 percentage points in multiclass and binary contexts respectively. The final noteworthy observation from the overall metrics is that, when moving from case 3 to 4, average precision improves by only 0.24 percentage points which is much less than expected based on other performance improvement trends that can be seen in table 18.

The much more surprising metrics are the class-wise statistics of which precision can be seen in table 19. In the context of metal class precision the multiclass cases dominate over the binary cases and the best metal precision can be found counterintuitively in the multiclass item-based case. The improvements in precision across the cases seem to occur in cardboard and non-metal and not across metal. Given that cardboard and plastic samples are more similar to each other it might be that random cases might bleed information about cardboard and plastic test sets allowing for overfitting. The most significant improvement can be found in the non-metal column when moving from the multiclass cases to the binary ones. The clear trade-off in terms of precision is that the multiclass models are extremely precise when classifying metal samples, but to attain that precision they struggle significantly with non-metal precision. Binary models on the other hand are much more precise with the non-metal samples but take a hit with metal sample precision.

Table 19. Class-wise cross-validation precision across cases.
The case 1 and 2 non-metal values are the average of the individual cardboard and plastic precisions*.

Case	Metal	Cardboard	Plastic	Non-metal
1	99.16%	76.38%	75.55%	75.97%*
2	98.64%	93.97%	74.75%	84.36%*
3	88.08%	-	-	97.46%
4	88.75%	-	-	99.71%

For class-wise recall the metal class shows incremental improvement when moving from item-based to random cases. However, the difference between the binary and multiclass item-based recall is very small whilst the difference between the random cases is quite a bit larger. For plastic, a

significant improvement in recall can be observed when moving from case 1 to case 2 while the corresponding cardboard recall deteriorates. This mirrors the pattern found in the class-wise precision table but the changes are much larger here. When referencing the case 2 confusion matrix in chapter 4, it can be observed that in this case the model heavily over-assigns plastic labels such that a large number of cardboard samples are classified as plastic, which explains these metrics. In terms of recall it is much more difficult to make an argument for using multiclass models over the binary ones as the binary models outperform the multiclass models in terms of recall albeit only slightly in the item-based case and moderately in the random case. Complete class-wise recall metrics are shown in table 20.

Table 20. Class-wise cross-validation recall across cases.
The case 1 and 2 non-metal values are the average of the individual cardboard and plastic recalls*.

Case	Metal	Cardboard	Plastic	Non-metal
1	94.63%	71.16%	83.99%	77.58%*
2	97.32%	64.28%	98.67%	81.48%*
3	95.17%	-	-	93.50%
4	99.46%	-	-	93.64%

Finally, the class-wise F1-score shown in table 21. It ties together the idea that if one is only interested in a good metal class performance, the ideal choice is a multiclass model instead of a binary model. Thus, the main benefit of using a binary model is the much higher performance with the non-metal classes. Though an argument could be made that due to the superior recall metrics demonstrated by binary models, one would still be more interested in using them in some waste classification contexts. In situations where positive class recall is much more relevant than balanced F1-metrics, like separating incinerable waste from non-incinerable waste, a binary model would likely be of more interest.

Table 21. Class-wise cross-validation F1-score across cases.
The case 1 and 2 non-metal values are the average of the individual cardboard and plastic F1-scores*.

Case	Metal	Cardboard	Plastic	Non-metal
1	0.9684	0.7368	0.7955	0.7662*
2	0.9798	0.7634	0.8506	0.8070*
3	0.9149	-	-	0.9544
4	0.9380	-	-	0.9658

It is worthwhile to note that there is a risk that the performance metrics for binary cases are somewhat biased towards the non-metal class. A demonstrable imbalance does exist between the number of samples in each class and while this imbalance is not extreme (a roughly 2:1 ratio of non-metal to metal samples) it can cause the model to over-prioritize correctly classifying non-metal samples over metal ones.

6 Conclusion

The purpose of this thesis was to examine the viability of microwave reflections as a data source for ML purposes in the field of material classification. The first step to examine this question was to evaluate the suitability of the measurement device for this task. This was put more formally in research question 1 as:

1. Can a vector network analyser be used to detect reflected microwave properties of plastic, cardboard and metal?

The answer to this question was examined in Chapter 4.4 data exploration. The visualisations presented show that the VNA is clearly able to detect varying power from metal samples. These reflections seem to be captured better in the transmitting antenna than the receiving antenna as the S11 parameter values show much wider variation than the S21 parameter values. This might be caused by the generated signal strength being too low to be registered by the receiving antenna, or that the receiving antenna was placed too far in the measurement system. The same visualisations also show that for most of the plastic and cardboard samples there is no perceivable variation in measurement results when objects are rotated. To ascertain which types of cardboard samples displayed reflective properties, the dataset would have to be re-measured such that tracking which measurements came from which item would also be included.

1. How well can an MLP network classify objects based on the object's reflected microwave properties?

The multiclass models were able to differentiate metals samples in the dataset with very good performance, and it was surprising that multiclass models were able to differentiate between the plastic and cardboard samples in the first place. Assuming that the multiclass models would have had 50% accuracy on plastic and cardboard samples and a 100% accuracy on the metal samples the expected accuracy would have been roughly 66% rather than the approximately 80% average accuracy that the multiclass models demonstrated. This implies that plastic and cardboard classification accuracy was not left to random chance and that some data-based decisions could be made by the models. However, it might be that the models are still incapable of generalizing cardboard and plastic classification beyond this dataset. Due to the dataset being so small, the DL models might have learned noise patterns that happen to coexist across samples of the same class instead of learning observable differences that can be used to separate objects of these classes, a classic case of overfitting. In order to have a more reliable assessment of model performance across

cardboard and plastic classes, more training data would be required in addition to real time classification testing.

The binary models had much better performance on the non-metal class but sacrificed performance on the metal class for it. Though it is very likely that the most important performance metric for a waste incineration facility is recall on the non-incinerable classes, i.e. metal. In this hypothetical the best overall model would be to a binary classification model.

6.1 Future Work

The potential for future work based on this experiment are vast. The investigation that was conducted in this thesis was intentionally very limited to retain a reasonable scope and not to overcommit resources on unproven and hypothetical systems. As such a significant number of potential avenues of testing were left out initially and more have sprung up from the findings in this experiment. While outlining possible future work and improvements suggestions, three different types of adjustment were deemed interesting, those being data measurement related, data modification related and real time testing.

6.1.1 Data Measurement Adjustments

The most significant improvement to this experiment would be to test with a wider signal bandwidth. Given that dielectric properties change depending on frequency, the sweep bandwidth of only 16 MHz is likely too small to capture frequency-dependent behaviour across multiple different material types. A larger measurement bandwidth could facilitate the system to find reflective qualities and rotational variation from materials other than metal, improving the potential use cases of the system significantly.

The second point of improvement would be to experiment with different antennae, mainly ones with sharper directionality. This would allow the measurement setup to be changed such that the Rx antenna could be placed closer to the Tx antenna without risking direct signal bleeding into the Rx antenna. As a consequence the received signal strength in the Rx antenna would likely increase and allow for more significant power deltas to be observed through the Rx antenna. Another way to increase the received signal strength would be to test with devices that allow for a more powerful microwave signal to be transmitted. In practice this would mean more powerful antennae and/or a more powerful signal generator. The signal strength generated by the nanoVNA was not examined

because no other alternatives were available but for future work examining signal strength is also a viable option.

The last immediately apparent adjustment would be to change the alignment of the Tx and Rx antennae. In this experiment they were placed roughly in parallel to test reflective traits, but given that reflection seems to be possible to examine using only the Tx antenna, the Rx antenna could be placed, for example, on the opposite side of the measured object in order to test absorptive behaviour. Another option would be to have multiple Tx and Rx antennae to measure the object from different angles and configurations simultaneously. This multiple antenna system does risk signal bleeding if not implemented with care.

6.1.2 Data Modifications

The way data was fed to the ML models in this thesis was very plain and as was shown by [5], experimenting with different data representations could increase model performance or otherwise give insights about the data. Possible adjustments could include, but are not limited to, testing with only S11 or S21 values, withholding the real or complex part of either S11 or S21 and adding vector magnitude as an additional representation.

The tested waste classes only contain very typical household waste and could easily be expanded to cover a much wider range of material classes like glass, wood, construction waste and so on. In general, any materials that could reasonably be expected to be found in a facility that handles waste material would be reasonable inclusions. Additionally, to properly test the system in a real-world environment it would have to be able to also classify the situation where no item is placed for measurement. This “null” object classification could be included to further validate the occurrence of microwave interaction, or lack thereof, in the plastic and cardboard item cases.

A limitation of the dataset is that it lacks distance variation. While this was intentionally left out, it means that, based on this experiment, no comments can be made about the relationship of distance to measured values other than what is shown in the formula for EPD (formula 23). By incorporating distance variation in follow-up work, this relationship could also be explored. The additional benefit of including distance variation is that it will increase the quantity of data available in the dataset if all other traits are kept the same, which could lead to more robust performance estimates for the ML models.

6.1.3 Real-time Testing

Possibly the best way to examine whether or not the proposed system is able to generalize its classification performance would be to take the trained models out of the theoretical context of performance estimation and perform real-time and real-world classification of waste items. This would require other steps to be taken beforehand, mainly creating a rigid measurement platform where the antennae alignment doesn't get altered from measurement session to the next and deciding how the antennae are placed relative to one another.

Another issue which arises from moving the system to a real-world context is it's robustness to noise. The sweep time of the nanoVNA-V2 F device is roughly 0.5-1 seconds. This slow sweep time can cause issues with moving targets or when items are placed for measurement. On a conveyor belt the position of the measured item at the start of the sweep can be different to its position at the end of the sweep. Similarly, when placing items in setups like the one used here can also cause fluctuations in the initial sweep due to the moving object or when a part of the human body is intercepting the Tx signal.

6.2 Final Words

In this thesis the potential of microwave reflections were examined as a source of sensor data for waste classification purposes. The results clearly indicate that the proposed measurement system is capable of detecting metal objects and that this data is suitable for use in ML classification with reasonable performance. To test the performance of the solution in a real-world context and examine the method's potential in detecting other types of materials, more research is required.

References

[1] Kaza, Silpa; Yao, Lisa C.; Bhada-Tata, Perinaz; Van Woerden, Frank. 2018. “What a Waste 2.0: A Global Snapshot of Solid Waste Management to 2050”. Urban Development;. © World Bank. http://hdl.handle.net/10986/30317 License: CC BY 3.0 IGO. Accessed: 8.10.2025

[2] Abdallah, Mohamed, et al. “Artificial Intelligence Applications in Solid Waste Management: A Systematic Research Review.” Waste Management (Elmsford), vol. 109, 2020, pp. 231–46, https://doi.org/10.1016/j.wasman.2020.04.057. Accessed: 6.10.2025

[3] Das, Sushrut. Microwave Engineering. 2014. Oxford University Press.

[4] Kim Ho Yeap. Electromagnetic Fields and Waves. Edited by Kim Ho Yeap and Kazuhiro Hirasawa, IntechOpen, 2019. Accessed: 26.4.2025

[5] Thomas, Athul and M, Manoj and P P, Prasanth, Classification of Targets From Microwave Reflections By Employing Deep Neural Networks (July 30, 2022). 4th International Conference on Communication & Information Processing (ICCIP) 2022, Available at SSRN: https://ssrn.com/abstract=4293567 or http://dx.doi.org/10.2139/ssrn.4293567 Accessed: 8.5.2025

[6] Goodfellow, Ian, et al. Deep Learning. The MIT Press, 2016.

[7] Yi, Chan Jia, and Chong Fong Kim. “AI-Powered Waste Classification Using Convolutional Neural Networks (CNNs).” International Journal of Advanced Computer Science & Applications, vol. 15, no. 10, 2024, https://doi.org/10.14569/IJACSA.2024.0151009. Accessed: 15.10.2025

[8] Al-Mashhadani, Israa Badr. “Waste Material Classification Using Performance Evaluation of Deep Learning Models.” Journal of Intelligent Systems, vol. 32, no. 1, 2023, pp. 1–16, https://doi.org/10.1515/jisys-2023-0064. Accessed: 15.10.2025

[9] Kamal, Shyam, et al. “Multilayer Hybrid Deep-Learning Method for Waste Classification and Recycling.” Computational Intelligence and Neuroscience, edited by Uma Seeboonruang, vol. 2018, no. 2018, 2018, pp. 1–9, https://doi.org/10.1155/2018/5060857. Accessed: 15.10.2025

[10] McCulloch, W.S., Pitts, W. A logical calculus of the ideas immanent in nervous activity. Bulletin of Mathematical Biophysics 5, 115–133 (1943). https://doi.org/10.1007/BF02478259 Accessed: 12.10.2025

[11] Hinton, G. (2012). Neural networks for machine learning. Coursera, video lectures. Available: https://www.cs.toronto.edu/~hinton/coursera_lectures.html. Accessed: 26.10.2025

[12] Diederik P Kingma, Jimmy Ba. Jan. 30, 2017. Adam: A Method for Stochastic Optimization. arXiv. Available: https://doi.org/10.48550/arXiv.1412.6980. Accessed 26.10.2025

[13] Thomas Krubiel, Shahrzad Khaleghian. Aug. 6, 2017. Training of Deep Neural Networks based on Distance Measures using RMSProp. arXiv. Available: https://arxiv.org/abs/1708.01911. Accessed: 26.10.2025

[14] Margherita Grandini, Enrico Bagli, Giorgio Visani. Aug. 13, 2020. Metrics for Multi-Class Classification: an Overview. arXiv. Available: https://arxiv.org/abs/2008.05756. Accessed: 4.8.2025

[15] R. Zamorano Ulloa, Ma. Guadalupe Hernandez Santiago, and V. L. Villegas Rueda, 'The Interaction of Microwaves with Materials of Different Properties', Electromagnetic Fields and Waves Chapter 1. IntechOpen, May 15, 2019. doi: 10.5772/intechopen.83675. Accessed: 24.4.2025

[16] Feynman RP, Leighton RB, Sands M. The Feynman Lectures on Physics. Mainly Electromagnetism and Matter. Vol. II. The New Millennium Ed. Basic Books; 2010. 566 p. ISBN: 978-0-465-07998-8. Accessed: 24.7.2025

[17] Shoaib, N. (2017). General Introduction. In: Vector Network Analyzer (VNA) Measurements and Uncertainty Assessment. PoliTO Springer Series. Springer, Cham. https://doi-org.ezproxy.utu.fi/10.1007/978-3-319-44772-8_1. Accessed: 5.2.2025

[18] Siglent. 2025. Vector Network Analyzer. Available: https://siglentna.com/vector-network-analyzer/. Accessed: 10.11.2025

[19] Anritsu. 2025. Vector Network Analyzers. Available: https://www.anritsu.com/en-us/test-measurement/network-analyzer/vector-network-analyzers. Accessed: 10.11.2025

[20] NanoRFE. Feb. 2022. NanoVNA-V2 User Manual. Available: https://nanorfe.com/nanovna-v2-user-manual.html#__RefHeading___Toc3237_3240808850. Accessed: 6.8.2025

[21] Copper Mountain Technologies. Apr. 26, 2023. What are Calibration Standards? Available: https://coppermountaintech.com/what-are-calibration-standards/. Accessed: 15.10.2025

[22] Laird Technologies. 2009. S8658PL S8658PR Circular Polarity RFID Panel Antenna. Available: https://gatewayrfidstore.com/product_images/Antenna/laird_indoor_etsi_specs.pdf Accessed: 23.3.2025

[23] TRAFICOM. Jul. 7, 2025. Frequencies and license holders of public mobile networks. Available: https://www.traficom.fi/en/communications/radio-licences-and-frequencies/frequency-planning-and-use/frequencies-and-license. Accessed: 8.5.2025

[24] Mohammed J. Uddin, Mohammad L. Hakim, Ismail Hossain. Dec. 2019. High Gain and Low Return-Loss Dual-Band DGS Antenna Loaded with Stub-Slot Configuration for 5G Wireless Communication. ReserachGate. Available: https://www.researchgate.net/publication/347818619_High_Gain_and_Low_Return-Loss_Dual-Band_DGS_Antenna_Loaded_with_Stub-Slot_Configuration_for_5G_Wireless_Communication. Accessed: 7.11.2025

[25] Sysjoint. Feb. 28, 2025. NanoVNA-F_V2. Available: https://www.sysjoint.com/index.php?tpl=product_detail&pid=11&uid=17&id=65&sno=1&list=2&lang=en. Accessed 8.5.2025

[26] Rune B. Broberg. 2025. NanoVNA-Saver. Available: https://github.com/NanoVNA-Saver/nanovna-saver/releases. Accessed 8.3.2025

[27] Keysight. 2025. Touchstone Format. Available: https://docs.keysight.com/display/genesys2010/Touchstone+Format Accessed: 13.3.2025

[28] GNU. Jun. 29, 2007. GNU General Public License. Available: https://www.gnu.org/licenses/gpl-3.0.en.html Accessed: 13.3.2025

[29] Rachel Ward. May 26, 2022. Stochastic Gradient Descent: where optimization meets machine learning. Lecture. Publisher: Institute of Advanced Study. Available: https://www.ias.edu/video/stochastic-gradient-descent-where-optimization-meets-machine-learning Accessed: 6.11.2025

Appendices

Appendix 1. Number of Individual Scans per Object in the Dataset

Table 22. Number of scans taken from each object in each class.

Object	Metal	Plastic	Cardboard
1	14	10	10
2	14	7	17
3	7	11	13
4	10	17	22
5	11	14	15
6	7	10	10
7	15	10	9
8	12	11	14
9	10	10	11
10	12	8	16
11	10	15	5
12	10	11	12
13	10	11	11
14	12	15	11
15	13	11	15
16	9	11	20
17	12	11	13
18	12	8	12
19	11	10	11
20	12	13	10
21	11	10	9
22	12	13	10
23	12	16	8
24	13	15	9
25	6	7	15
26	11	10	15
27	12	12	10
28	8	14	14
29	12	11	12
30	8	11	6
31	10	16	-
32	12	9	-
33	11	7	-

Object	Metal	Plastic	Cardboard
34	12	-	-

Appendix 2. Remaining Object-wise S-parameter Visualisations

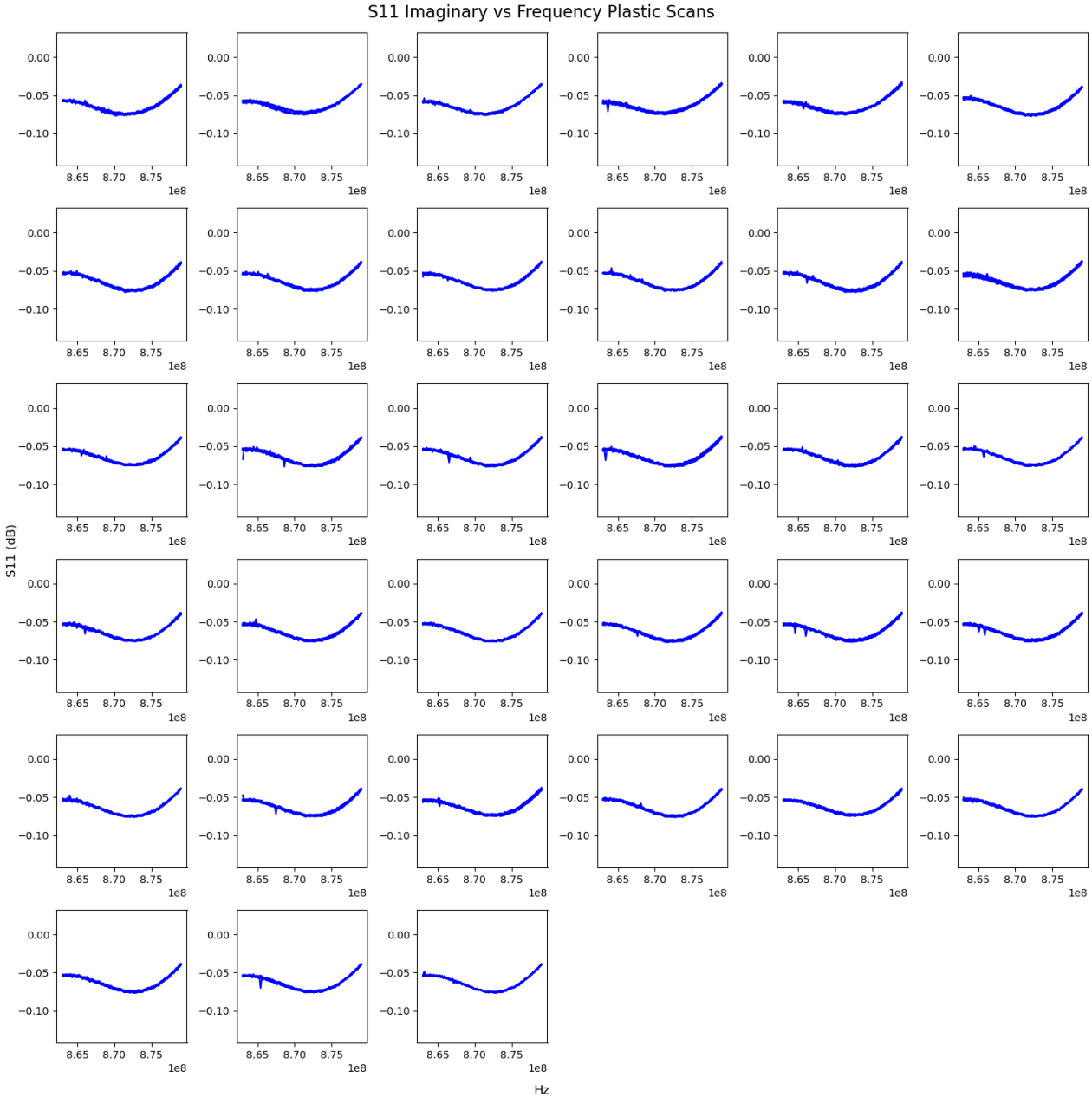

Figure 30. Imaginary part of S_{11} measurements of plastic samples by item.

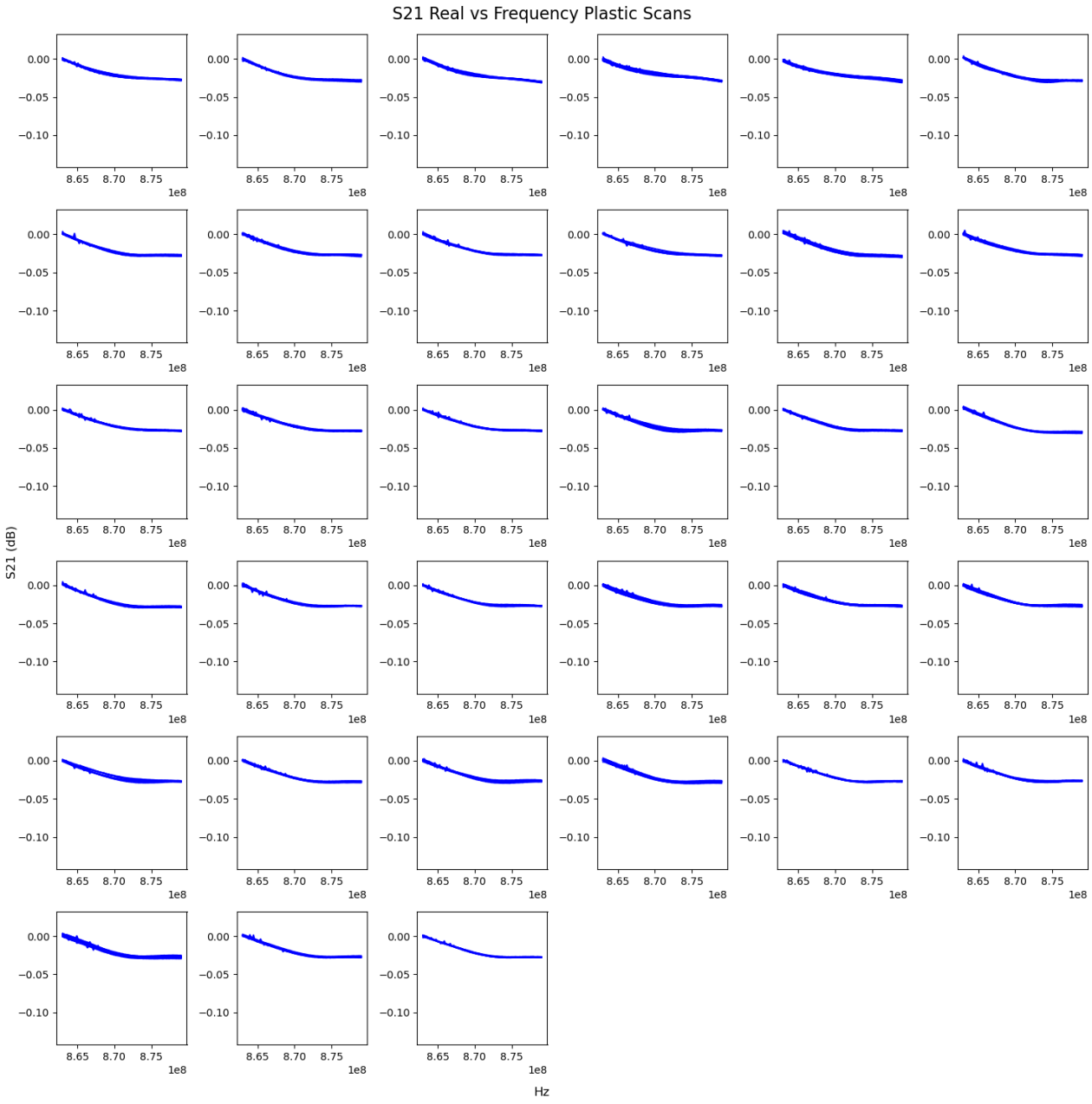

Figure 31. Real part of S_{21} measurements of plastic samples by item.

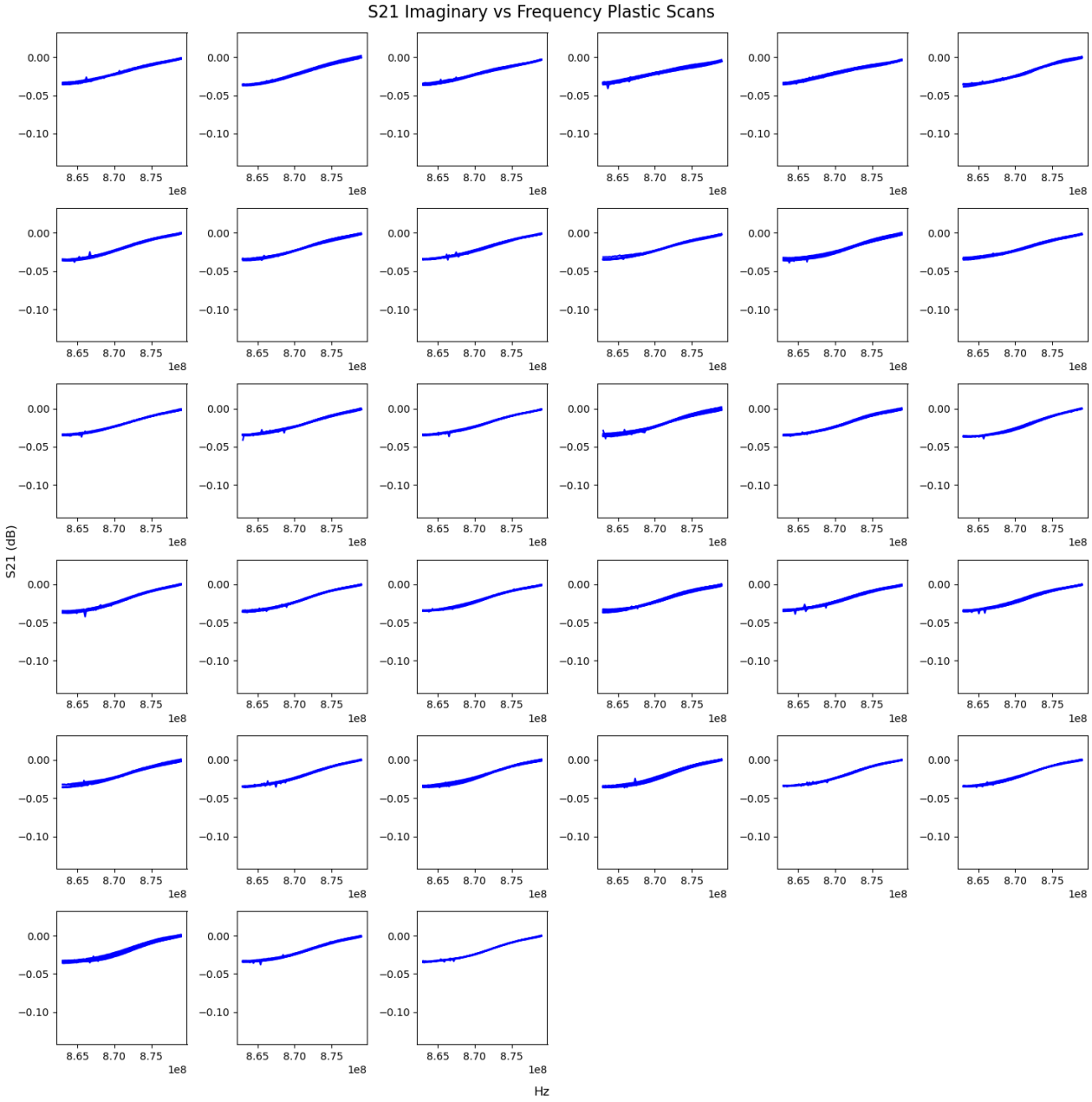

Figure 32. Imaginary part of S_{21} measurements of plastic samples by item.

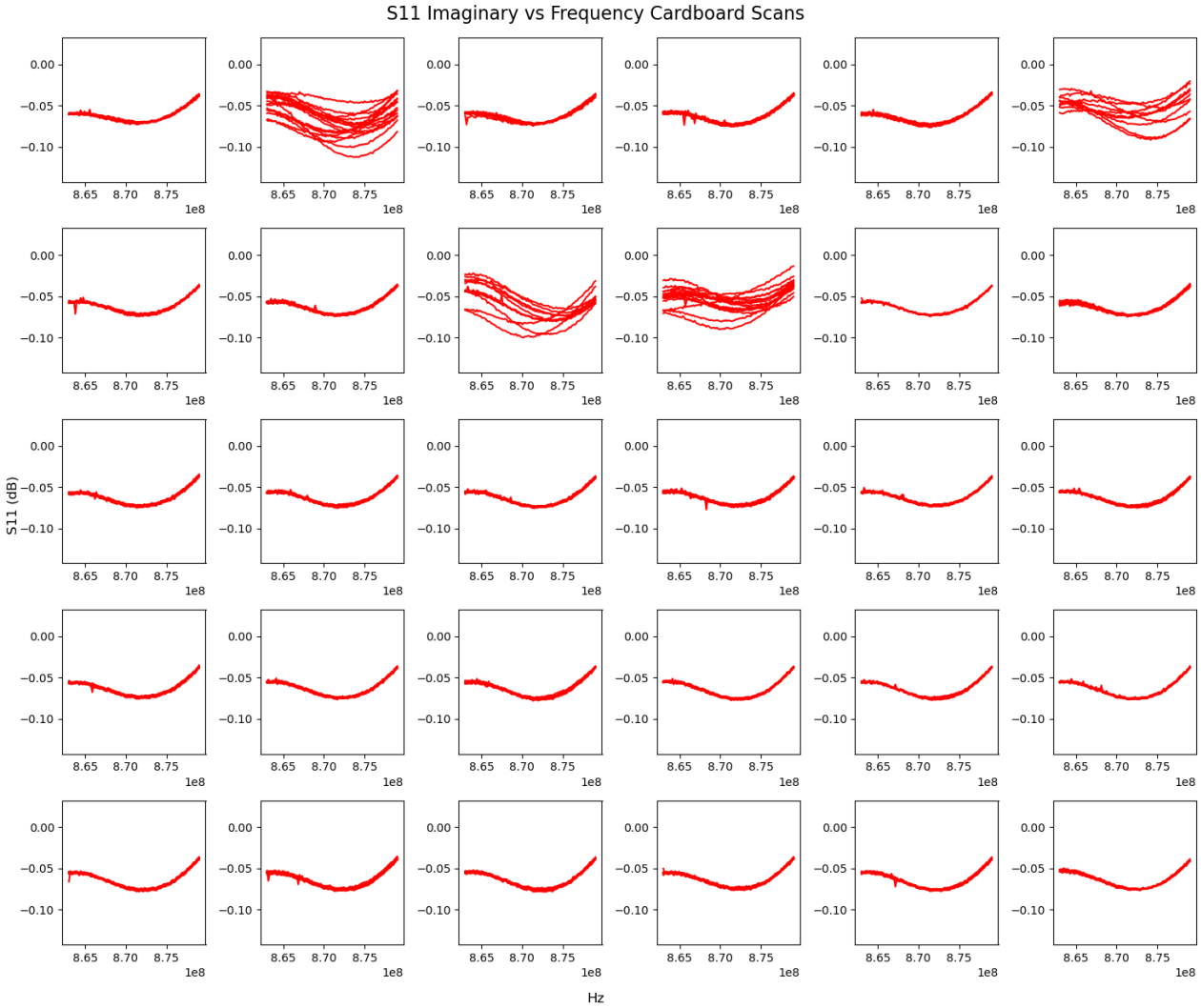

Figure 33. Imaginary part of S_{11} measurements of cardboard samples by item.

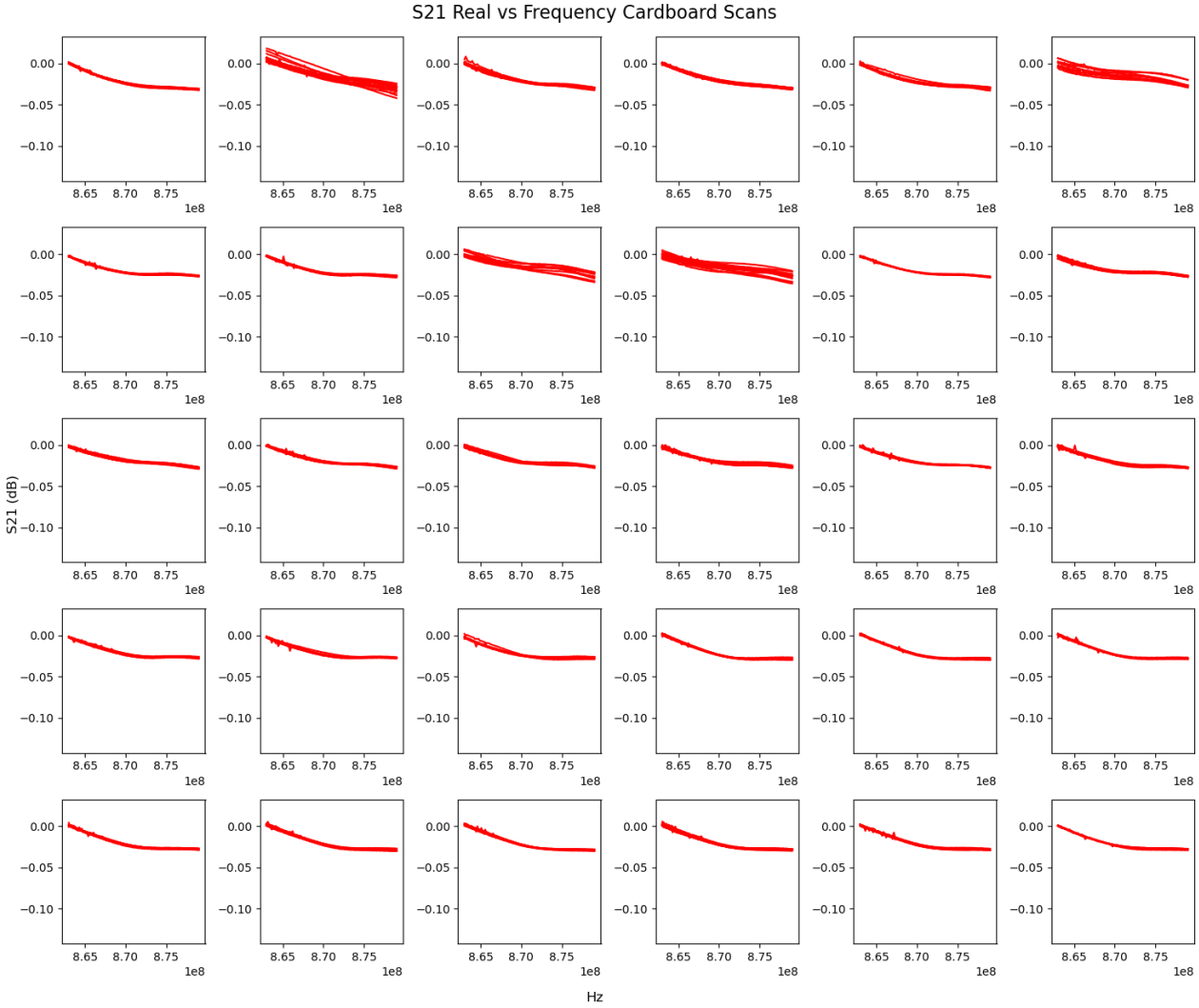

Figure 34. Real part of S_{21} measurements of cardboard samples by item.

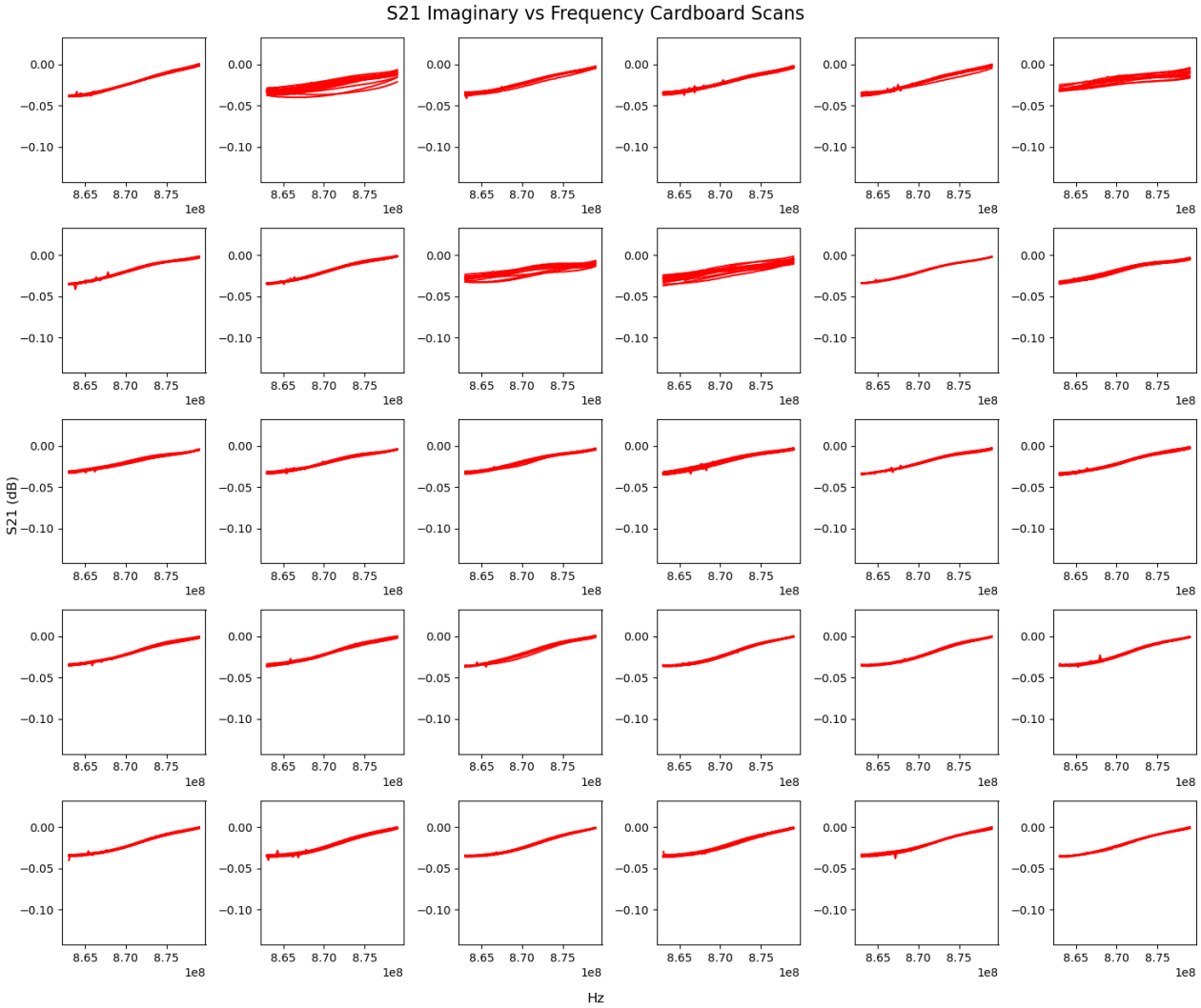

Figure 35. Imaginary part of S_{21} measurements of cardboard samples by item.

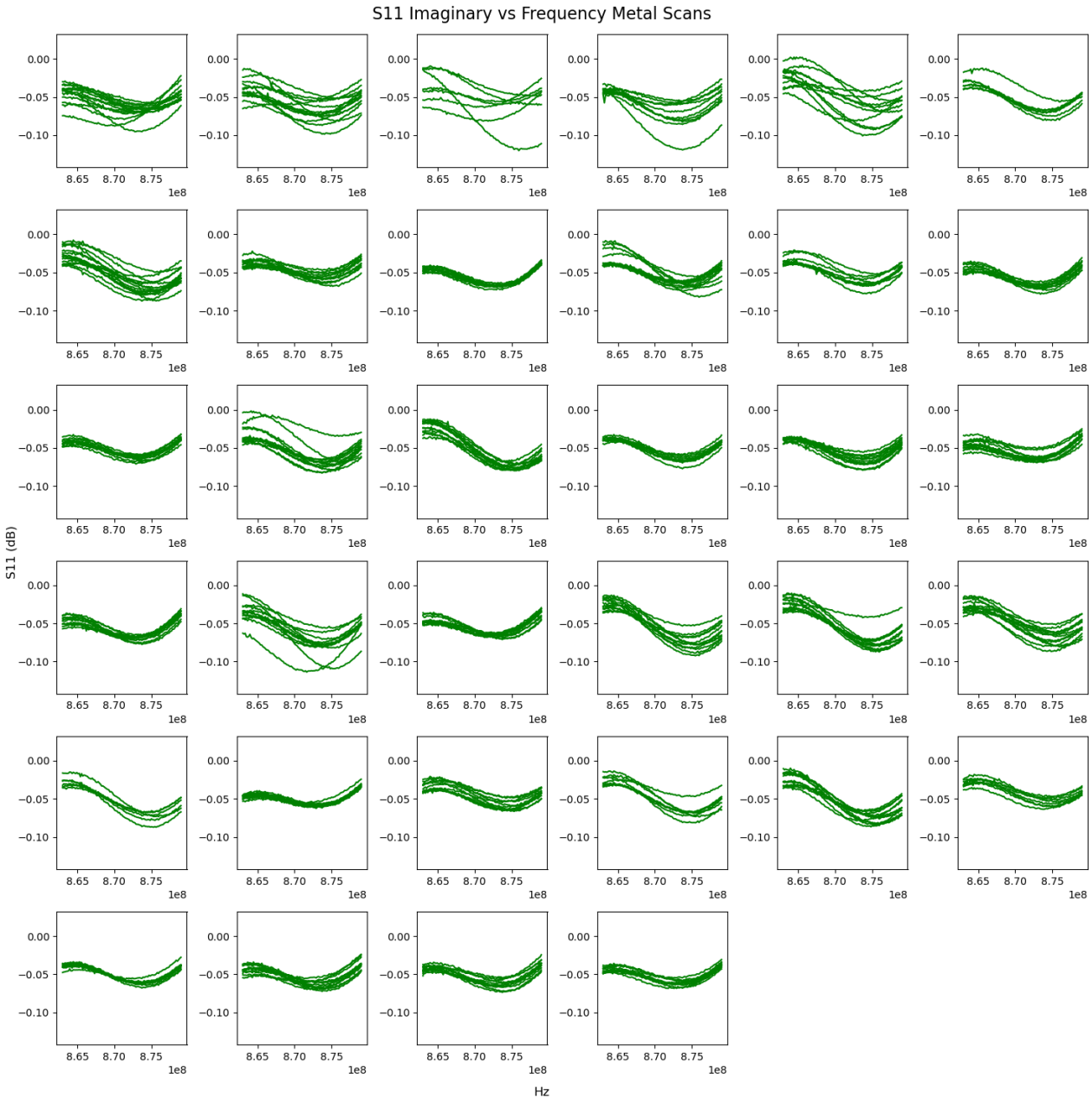

Figure 36. Imaginary part of S_{11} measurements of metal samples by item.

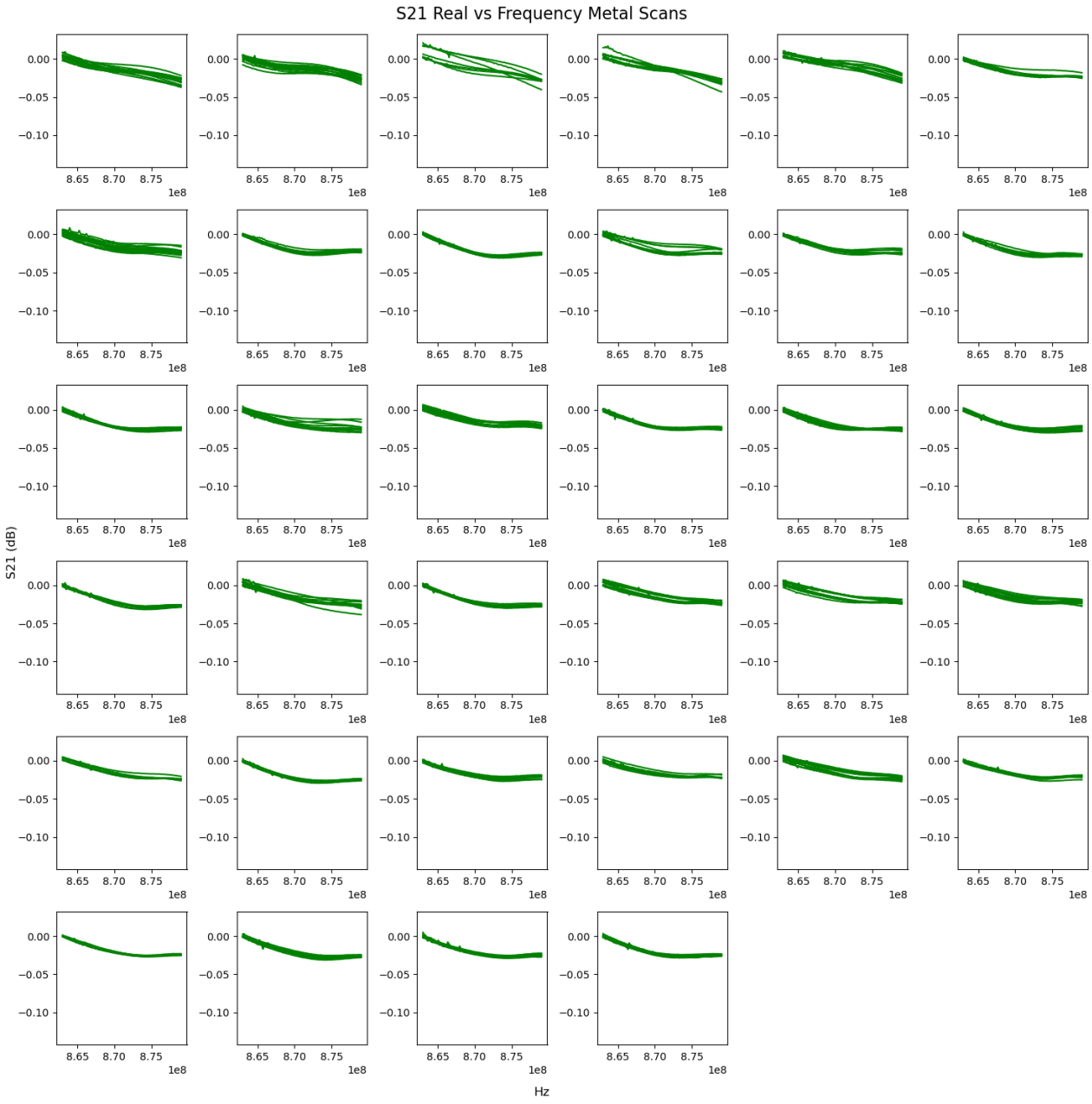

Figure 37. Real part of S_{21} measurements of metal samples by item.

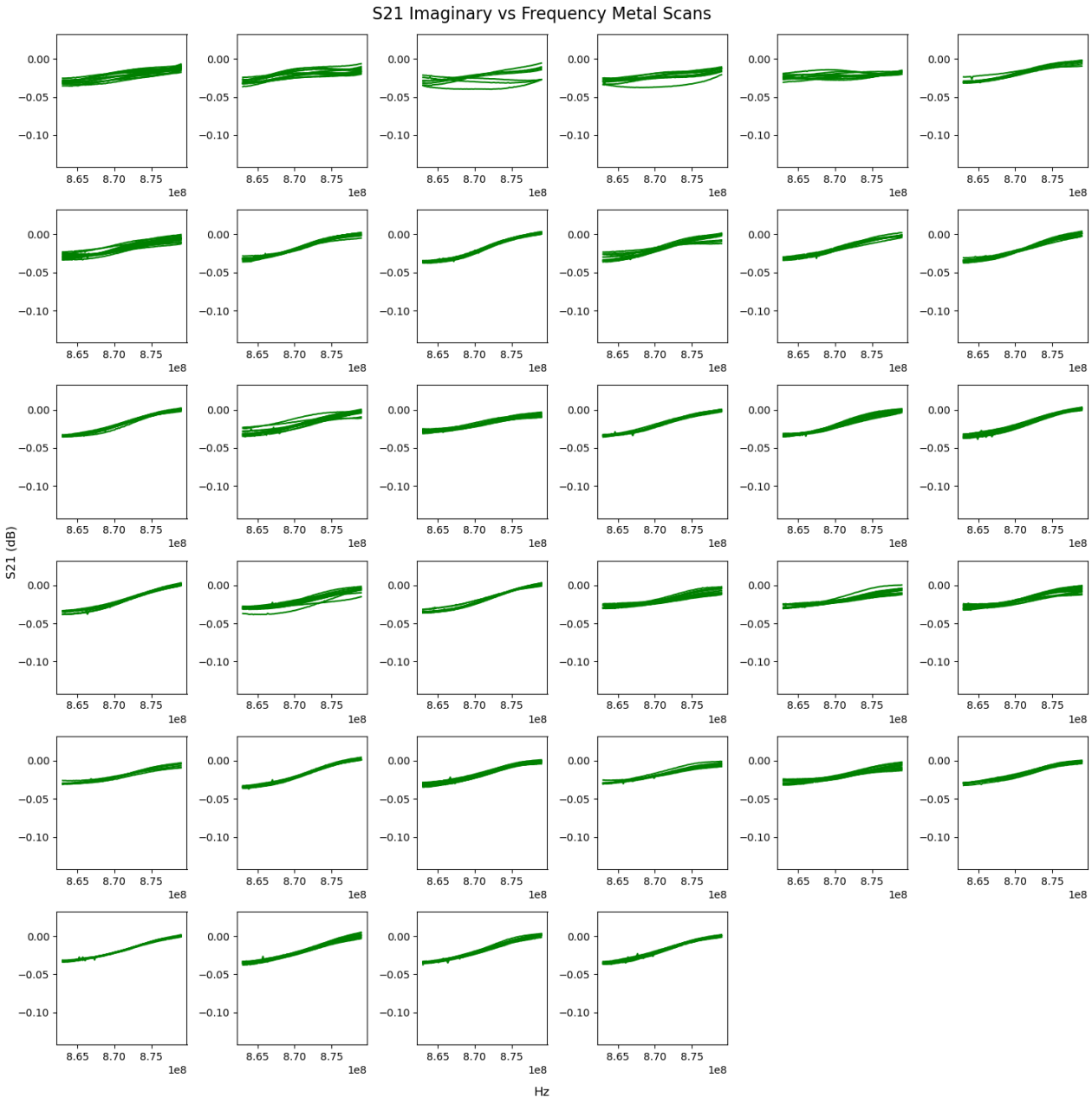

Figure 38. Imaginary part of S_{21} measurements of metal samples by item.