This article has been accepted for publication in IEEE Transactions on AgriFood Electronics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFE.2025.3624544

Can ML Enhance the Accuracy of Crop Quality Assessment with Radio Frequency Reflectometry?

Adeoluwa Oyinlola*, Temitope Odedeyi†
Dept. of Electronic and Electrical Engineering, University College London, London WC1E 6BT
*a.oyinlola@ucl.ac.uk, †t.odedeyi@ucl.ac.uk

***Abstract*—This paper presents a novel approach that integrates AI/ML algorithms-Linear regression, Random forest, and Principal Component Analysis (PCA) to enhance the predictive accuracy of Radio Frequency reflectometric (RF-R) analysis for crop quality estimation. Using 131 samples of cassava roots, RF-R measurements were carried out within 30 kHz to 200 MHz frequency range to estimate dry matter as a proxy for starch content. Multiple modeling strategies were evaluated, including iterative linear regression, random forest, and principal component analysis, which were applied to both single-frequency and multi-frequency combinations. A single-frequency analysis from simple linear regression achieved an R^2 of 0.64, while, for the first time, we show that combining two and three frequencies for a random forest model enhances prediction accuracy by 23.4%, highlighting the potential for improved crop quality assessment. Overall, this work demonstrates that combining RF-R with ML improves the specificity and robustness of quality prediction, establishing a foundation for application in real-time, non-destructive crop assessment.**

***Index Terms*—RF Reflectometry, Agricultural Phenotyping, Non-Destructive Testing, Machine Learning, Dry Matter Estimation**

I. Introduction

RF reflectometry (RF-R), a method of impedance spectroscopy, is a powerful and cost-effective technique for non-destructive materials characterisation [1]. With this technique, a sample under test is excited by an RF signal and the properties of the signal reflected to the source are used to infer material properties and composition [2], [3]. This method has been widely applied in diverse fields, particularly in the characterisation of food and agricultural products [4]. These RF-R responses are influenced by the physical and chemical composition of crops, including moisture content, starch concentration, and fibre structure, making the technique a promising tool for real-time, field-deployable quality assessments [5].

Alternative methods such as chemical analysis and Near-Infrared Spectroscopy (NIRS) are conventional approaches for starch content estimation and have achieved significant successes for evaluating crop and food quality attributes; however, they present notable trade-offs. These techniques can be time-consuming, destructive, and impractical for high-throughput or in-field use [6], [7]. Chemical analysis requires sophisticated laboratory setups and skilled personnel, while NIRS, despite being fast and non-destructive, can be expensive and sensitive to calibration and sample heterogeneity [8], [9].

However, RF-R analysis has limitations, notably its non-specificity — RF-R measurements provide an aggregate response influenced by multiple physical and chemical properties, making it challenging to isolate specific traits. In complex biological systems, where multiple interacting components influence electrical responses, accurately interpreting reflectometric data remains a challenge.

To address this, our study explores the integration of ML algorithms with RF-R to enhance prediction accuracy. The study is specifically focused on predicting starch content in cassava (*Manihot esculenta*)—a tropical root crop that serves as a staple for 800 million people and a key industrial raw material [10], [2]. Starch content is crucial in determining cassava's economic value and industrial applicability. By leveraging ML models trained on RF-R data, we aim to improve the specificity and robustness of this technique, making it a viable alternative for high-throughput crop quality assessment. Additionally, our findings pave the way for broader applications, including contamination detection and food adulteration analysis.

The remainder of this paper is structured as follows: Section II describes the experimental methodology, including sample preparation, RF measurements, and data collection. Section III presents the AI-based modeling approach and feature extraction techniques, covering both traditional regression methods and ML techniques such as Random Forest and PCA. Section IV discusses the results obtained from various modeling approaches, comparing the predictive performance of different frequency selection methods, as well as evaluating computational efficiency and the trade-offs between explainability and accuracy. Finally, Section V concludes the paper by summarizing key insights, highlighting the implications of these findings, and suggesting future directions for improving crop quality assessment using RF-R and ML-driven modeling.

Abbreviations

- **RF** - Radio Frequency
- **DM** - Dry Matter
- **RF-R** - RF Reflectometry
- **ISLR-1F** – Iterative Simple Linear Regression for 1 Frequency
- **IRF-1F** – Iterative Random Forest for 1 Frequency
- **IMLR-2F (All)** – Iterative Multiple Linear Regression for 2-Frequency Combinations
- **IMLR-3F (20k)** – Iterative Multiple Linear Regression for 3-Frequency Combinations (20,000 Sampled)
- **IMRF-2F (5k)** – Iterative Random Forest for 2-Frequency Combinations (5,000 Sampled)
- **IMRF-3F (5k)** – Iterative Random Forest for 3-Frequency Combinations (5,000 Sampled)

This article has been accepted for publication in IEEE Transactions on AgriFood Electronics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFE.2025.3624544

II. Materials and Method

A. Materials

A total of 60 cassava samples were obtained from local grocery stores. These samples, predominantly imported from Latin America, belonged to the sweet cassava variety. This variety is characterised by its lower cyanide content, approximately 14% sucrose, and trace amounts of fructose and dextrose [11]. To extend their shelf life, these roots had been coated in wax soon after harvest [12]. For this experiment, the effect of wax coating was not explored; however, the dielectric constant of paraffin wax, which is between 2 - 3.5 [13] will have negligible effects on the sample measurement relative to the dielectric constant of water, which is about 80 at the observed frequencies. The obtained samples were carefully stored and transported to the laboratory for measurement. Each sample was cut into latitudinal sections of approximately 5cm thickness, yielding 131 distinct observations.

B. Method

Following procedures similar to those reported in [2], [14], [5], RF-R measurements were carried out using a Keysight FieldFox Microwave Analyser N9917A, a vector network analyser (VNA) with a custom-designed dual-pin probe as shown in Fig. 1(c) to record the one-port reflection coefficient of the cassava samples across a frequency range of 30 kHz to 200 MHz. The relationship between return loss and RF-R coefficient is shown in (1) and (2). The custom-designed dual-pin probe replaced the open-ended coaxial probe used in early RF-R crop quality estimation experiments [15]. The slim-form structure of the open-ended coaxial probe resulted in limited surface contact and required multiple measurements especially for large or inhomogeneous samples, which disrupts sample integrity [14], [2], [16]. To overcome these limitations, a dual-pin probe was designed by connecting the coaxial conductors to two 5mm pins placed 15mm apart. This configuration increased contact areas, minimised sample disruption, and provided more consistent and reliable measurements [16].

The measurement setup is illustrated in Fig. 1(a). The probes were carefully inserted into the roots, penetrating through the peel and into the flesh.

$$\Gamma = \frac{Z_{SUT} - Z_o}{Z_{SUT} + Z_o} \tag{1}$$

$$RL = -20 \log_{10} |\Gamma| \tag{2}$$

where Z_{SUT} is impedance of the sample under test, Z_o is source characteristic impedance (which is usually 50 Ω in RF and microwave test setup), Γ is RF-R coefficient, and RL is Return Loss in dB.

To account for the potential inhomogeneity in the distribution of starch within the cassava roots, each section was measured at four distinct points around its surface and then averaged. This approach ensured that the recorded frequency spectrum was representative of the sample's overall composition, minimising measurement variability.

Following the measurement of RL, the samples were cut into smaller pieces and placed in paper bags and weighed using a Kern EMB 500-1BE digital scale (d = 0.1g). These values were recorded as the sample fresh weights. The samples were afterwards placed in a Heraeus UT 6060 laboratory oven for drying at an average temperature of $90°C$ until a constant weight was reached, indicating the complete removal of water. From [3], the DM content was calculated as the ratio of the dry weight to the fresh weight of the sample [16].

$$\text{Dry Matter } (\%) = \frac{\text{Dry Weight}}{\text{Fresh Weight}} \times 100 \tag{3}$$

The spectra plot of the magnitude of all reflection coefficient values across all sampled frequencies were then plotted for each measured sample, as shown in Fig. 1(b).

C. Experimental Objectives

The aforementioned modifications introduced changes to the frequency response of the probe-sample system compared to previous experimental setups [16]. Given this shift, the study aimed to systematically assess frequency selection for optimal DM estimation by identifying the optimal frequency for DM using the dataset, investigating multi-frequency selection strategies to determine whether a combination of two or more frequencies could improve prediction accuracy and reduce the MSE and lastly, develop and evaluate a predictive model capable of leveraging the optimal frequency selection to enhance the robustness and precision of cassava DM content estimation.

III. Analysis and Modelling

AI models, including various ML algorithms such as random forest, may offer the potential to uncover deeper correlations between dielectric properties and DM content. By leveraging ML-driven feature selection and regression techniques, it may be possible to achieve more precise and generalisable predictions, further improving the reliability of cassava quality assessment [17].

A. Single Frequency Selection

1) Iterative Simple Linear Regression (ISLR): To identify the most relevant frequency for predicting DM content, an iterative simple linear regression approach was employed across all 401 frequency-dependent features. The objective was to determine the frequency that maximised the coefficient of determination (R^2) while minimizing the MSE, ensuring the best possible linear fit to the target variable.

Simple linear regression is based on modeling the relationship between an independent variable X and a dependent variable Y using a linear equation [18] :

$$Y = \beta_0 + \beta_1 X + \epsilon \tag{4}$$

where:

- Y represents DM (%),
- X is the selected frequency,
- β_0 is the intercept, representing the predicted DM content when $X = 0$,

This article has been accepted for publication in IEEE Transactions on AgriFood Electronics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFE.2025.3624544

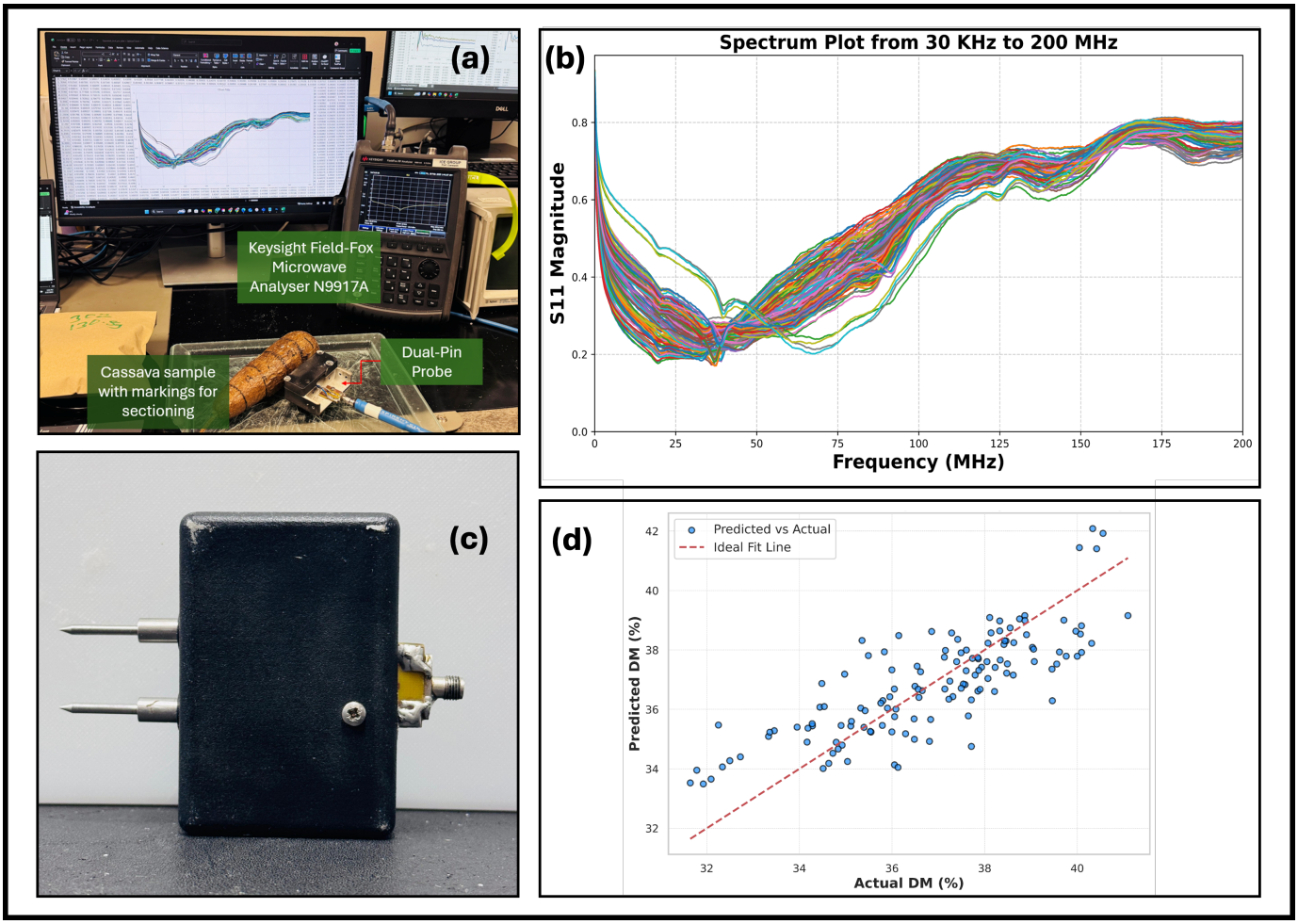

Fig. 1. **(a)** Experimental setup showing the Keysight FieldFox Microwave Analyser (N9917A) connected to a dual-pin probe for RF-R measurements on cassava samples prepared with markings for sectioning. **(b)** Representative S_{11} magnitude spectra of cassava samples measured between 30 kHz and 200 MHz, illustrating frequency-dependent variations across samples. **(c)** Close-up view of the custom dual-pin probe. **(d)** Regression plot of predicted versus actual DM % using the best single-frequency selection, 18.5 MHz with an R^2 of 0.64 and MSE of 1.65

- β_1 is the regression coefficient, indicating the rate of change in DM (%) per unit increase in the frequency, and
- ϵ is the error term, accounting for variations in DM content not explained by the linear model.

The R^2 statistic is defined as:

$$R^2 = 1 - \frac{\sum(Y_i - \hat{Y}_i)^2}{\sum(Y_i - \bar{Y})^2} \tag{5}$$

where:

- Y_i represents the actual observed DM (%) values,
- $\hat{Y}_i$ denotes the predicted values obtained from the regression model, and
- $\bar{Y}$ is the mean of the observed DM (%) values.

The numerator of the fraction represents the residual sum of squares (RSS), which quantifies the error between the predicted and actual values, while the denominator represents the total sum of squares (TSS), capturing the total variation in DM (%). A higher R^2 value signifies that a larger proportion of variance in DM content is explained by the frequency, indicating a stronger predictive relationship.

To achieve these objectives, we applied simple linear regression, fitting a line of best fit for individual frequency against DM (%). This method identified 18.5 MHz as the new optimal frequency with an R^2 score of 0.6375, regression plot shown in Fig. 1(d), replacing the previously established 30 MHz, which had an R^2 of 0.5414 [16]. This shift in coefficient of determination from the 30 MHz frequency to 18.5 MHz frequency is attributed to changes in probe design as well as experimental method as described in IIB. Statistical metrics were used to verify the biological significance and relationship of the selected frequency with DM. The t-value was 15.06 and the corresponding p-value was 3.3×10^{-30}, indicating that 18.5 MHz was highly significant. This result confirms that the selected frequency contributes meaningfully to explaining the observed variation in DM, reinforcing its potential as a reliable predictor in non-destructive field-based estimation.

2) Frequency Clustering Pattern: Using iterative simple linear regression, we systematically analysed the predictive power of each frequency in relation to DM (%), selecting the top 50 frequencies that yielded the highest R^2 values. Notably, these frequencies were not randomly distributed across the spectrum but instead exhibited distinct clustering patterns,

indicating specific frequency bands with stronger predictive relationships to DM content. The first cluster emerged within the range of 3.5 MHz to 10 MHz, where the predictive strength was characterised by an R^2 score of 0.62. A second and even more prominent cluster was observed between 14 MHz and 23.5 MHz, achieving the highest R^2 score of 0.64, suggesting a relatively stronger association between DM content and dielectric properties in this frequency range for the observed dataset. Finally, a third cluster was identified within the 106 MHz to 110 MHz range, where the predictive performance remained substantial with an R^2 score of 0.60. The emergence of these clusters suggests that certain frequency ranges may be inherently more sensitive to variations in DM content, potentially due to dielectric behaviour, or interactions between electromagnetic waves and the internal structure of the cassava samples [19], [20].

3) Iterative Random Forest (IRF): Random forest was employed as an algorithm to evaluate the predictive power of various frequency features on the target variable, DM. Random forest, an ensemble learning method, was utilised to assess the relationship between each frequency and DM by training and testing models on a series of frequencies independently [21], [22].

To assess robustness, model training and evaluation were repeated across randomised splits, and performance metrics (R^2, MSE) were reported in TABLE I. Initially, the dataset was split into training (70%) and testing (30%) subsets to ensure a robust evaluation of model performance. This stratified split helps in reducing overfitting and provides a reliable estimate of how the model will perform on unseen data [23]. The dataset was then subjected to a feature-wise analysis, where each frequency was considered independently to predict DM [24].

The random forest model was initialised with 100 decision trees (n_estimators = 100) and trained for each frequency in the dataset. The predictions of these individual trees are then aggregated, typically by averaging, to produce a final prediction [25], [26]. This ensemble approach improves the model's generalisation ability, reduces overfitting, and captures complex, non-linear relationships between features and the target variable.

The model's performance was evaluated using R^2, which quantifies the proportion of variance in DM explained by the model. Higher R^2 values indicate better predictive performance. After computing the R^2 score for each frequency, the frequency with the highest R^2 value was identified as the most predictive feature for DM.

The result identifies 13.5 MHz with the highest R^2 score while demonstrating a substantial predictive capability, with an R^2 value of 0.68. This indicates that the model, using this frequency alone, explains approximately 68% of the variance in DM, which is a strong indicator of the frequency's importance in predicting DM. The corresponding t-value of 14.87 and extremely low p-value of 9.3×10^{-30} further confirm the significance of this relationship, reinforcing the frequency's relevance in modeling DM.

It is important to note that the use of random forest in this exploration does not necessarily mean that the identified frequencies are the most biologically relevant for estimating DM (%); rather, they are merely the optimal frequencies at training the random forest model to make more generalisable predictions.

B. Multiple Frequency Combinations

1) Assessing Multicollinearity In Multiple Feature Combinations: Before combining multiple frequencies for DM prediction, it is essential to evaluate multicollinearity among frequency features. Multicollinearity occurs when independent variables (frequencies) are highly correlated with each other, which can distort regression coefficients and reduce the interpretability of predictive models [27]. To address this, we employed stepwise feature selection combined with Variance Inflation Factor (VIF) filtering, ensuring that selected frequencies are not only strongly correlated with DM but also exhibit minimal redundancy with other frequencies. A high VIF value (typically above 10) indicates strong collinearity, suggesting that the variable may be redundant [28], [29].

A threshold of $|0.5|$ was established for correlation with DM, meaning that a frequency had to be at least moderately correlated with DM, while also maintaining a low VIF score. Through this process, 18.5 MHz was identified as the singular frequency that met both criteria: it demonstrated the highest individual correlation with DM and exhibited minimal collinearity with other frequencies.

This finding is critical for experimental design and future studies seeking to optimise DM prediction. Given that 18.5 MHz achieved an R^2 score of 0.64, this indicates that it accounts for 64% of the variability in DM content. If additional frequencies are incorporated into the predictive model, they would contribute at most 36% to the remaining explainability, making their selection crucial for model improvement.

To identify supplementary frequencies that could enhance prediction accuracy, we assessed:

- The correlation of each frequency with DM.
- The correlation of each frequency with 18.5 MHz.

From this analysis, 235 frequencies were identified that exhibited strong correlation with both DM and 18.5 MHz. To determine their predictive contributions, multiple linear regression analyses were conducted, using 18.5 MHz as the reference frequency in combination with each of the identified frequencies. The resulting R^2 values ranged from 0.6375 to 0.673, with the highest predictive performance observed when combining 18.5 MHz and 20.5 MHz, yielding an R^2 score of 0.673.

These results suggest that while 18.5 MHz is the dominant predictor, carefully selected secondary frequencies can enhance DM prediction. This underscores the importance of systematically evaluating multicollinearity before integrating multiple frequencies in predictive modeling.

2) Iterative Multiple Linear Regression for Two-frequency Combination (IMLR-2F): In this study, we sought to identify the most predictive frequency pairs for estimating the DM content in cassava tubers using RF reflectometry. To achieve this, an exhaustive evaluation was conducted on all possible two-frequency combinations from a dataset containing 401

This article has been accepted for publication in IEEE Transactions on AgriFood Electronics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFE.2025.3624544

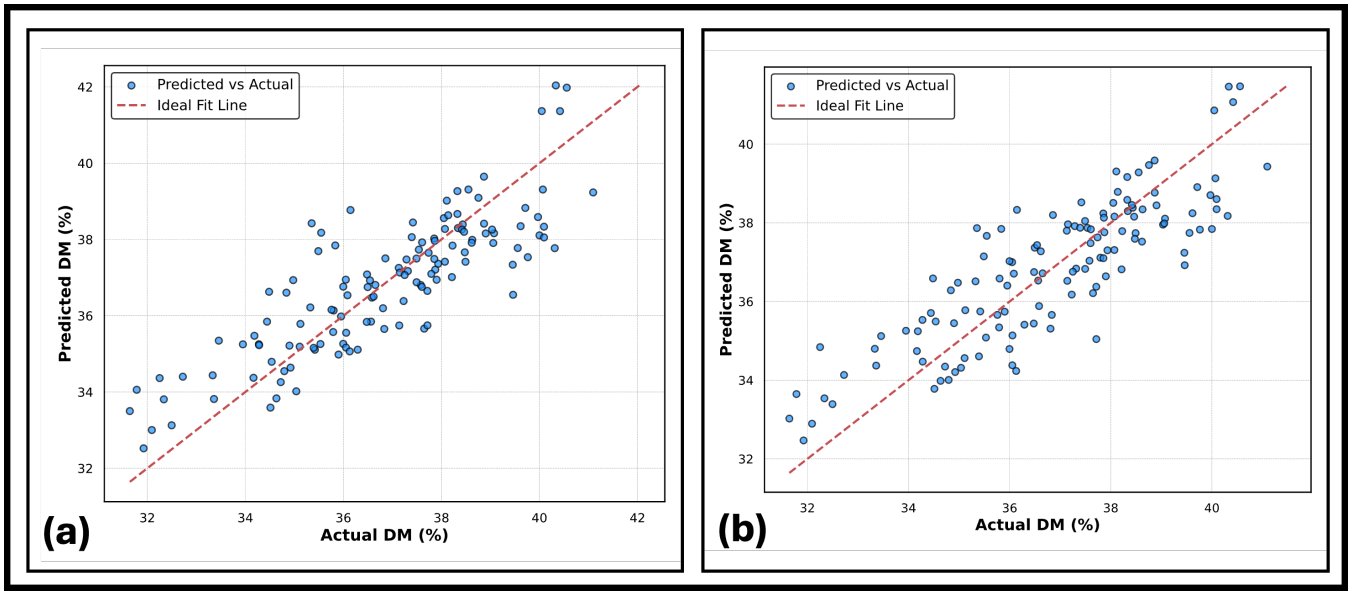

Fig. 2. Predicted vs. actual DM content using iterative multiple linear regression. **(a)** regression with features at 0.5 MHz and 1.0 MHz achieving $R^2 = 0.707$ and $\mathrm{MSE} = 1.33$. **(b)** regression with features at 0.5 MHz, 3.5 MHz, and 80 MHz, achieving $R^2 = 0.720$ and $\mathrm{MSE} = 1.28$.

frequency features, each corresponding to a specific RF measurement taken during the reflectometry process.

The analysis was performed by applying multiple linear regression to each pair of frequencies, where the independent variables were the two selected frequency features, and the dependent variable was the DM % content. The linear regression model was iteratively fitted to all 80,200 unique frequency pairs, which were generated using the *itertools.combinations* method [30]. For each combination, the model's performance was evaluated using the R^2 score and MSE, metrics that measure how well the two-frequency combination explained the variance in the DM % content. After fitting the linear regression models and calculating the R^2 scores for each, the results were ranked to identify the top-performing frequency combinations.

One of the top-performing frequency pairs was 500 kHz and 1.0 MHz, which yielded an R^2 score of 0.71 and MSE of 1.33. The regression plot is shown in Fig. 2(a). This indicates that the combination of these two frequencies explained approximately 71% of the variance in the DM % content, suggesting a significant predictive relationship. The model's overall significance was confirmed by an F-statistic of 154.78 and a p-value of 6.84×10^{-35}, indicating that the combined use of both frequencies captures substantial variation in DM and enhances the robustness of prediction. It is important to note, however, that the identified frequency pairs may shift slightly when the data is reshuffled and the iteration is run again. The performance of different frequency combinations may vary with different random samplings of the data, but the overall trend of identifying the most predictive frequencies remains consistent.

3) Iterative Multiple Linear Regression for Three-frequency Combinations (IMLR-3F): In this analysis, the goal was to identify the best three-frequency combinations for predicting DM% in cassava tubers using linear regression. Similar to the two-frequency combination analysis, this process involved evaluating a subset of possible frequency combinations rather than examining every combination exhaustively. Hence, from the 10,000,660 possible combinations, a random sample of 20,000 combinations was selected for evaluation. The approach was designed to balance the need for computational efficiency with the goal of identifying the most effective frequency combinations for training the model.

Analysis of three-frequency combinations for predicting DM% in cassava tubers revealed that the combination 0.5 MHz, 3.5 MHz and 80 MHz produced the highest R^2 score of 0.72. This outcome, however, is not definitive. Given the stochastic nature of the random sampling and the iterative nature of the modeling process, it is possible that further iterations or reshuffling of the data could yield different results. It is interesting to note that the range of the RF-R coefficient up to 5 MHz (see Fig. 2(b)) is relatively narrow compared to higher frequencies across all observed samples. This suggests that selecting 0.5 MHz and 3.5 MHz may result in increased susceptibility to measurement-induced noise. In fact, additional iterations or alternative combinations of frequencies could potentially lead to a combination that produces an even higher R^2 score. However, the model's validity was supported by an F-statistic of 108.59 and a p-value of 6.75×10^{-35}, confirming that the combination of one or more of the three frequencies significantly contributes to explaining variation in DM and strengthens model generalisability.

4) Iterative Random Forest modeling for two-frequency combinations (IRF-2F): This modeling approach aimed to identify two-frequency combinations that enable the random forest model to achieve the highest predictive performance for DM (%).

To minimise computational overhead, 5,000 frequency pairs were randomly selected and modeled from the 80,200 possible two-frequency pairs. For each selected combination, a random

This article has been accepted for publication in IEEE Transactions on AgriFood Electronics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFE.2025.3624544

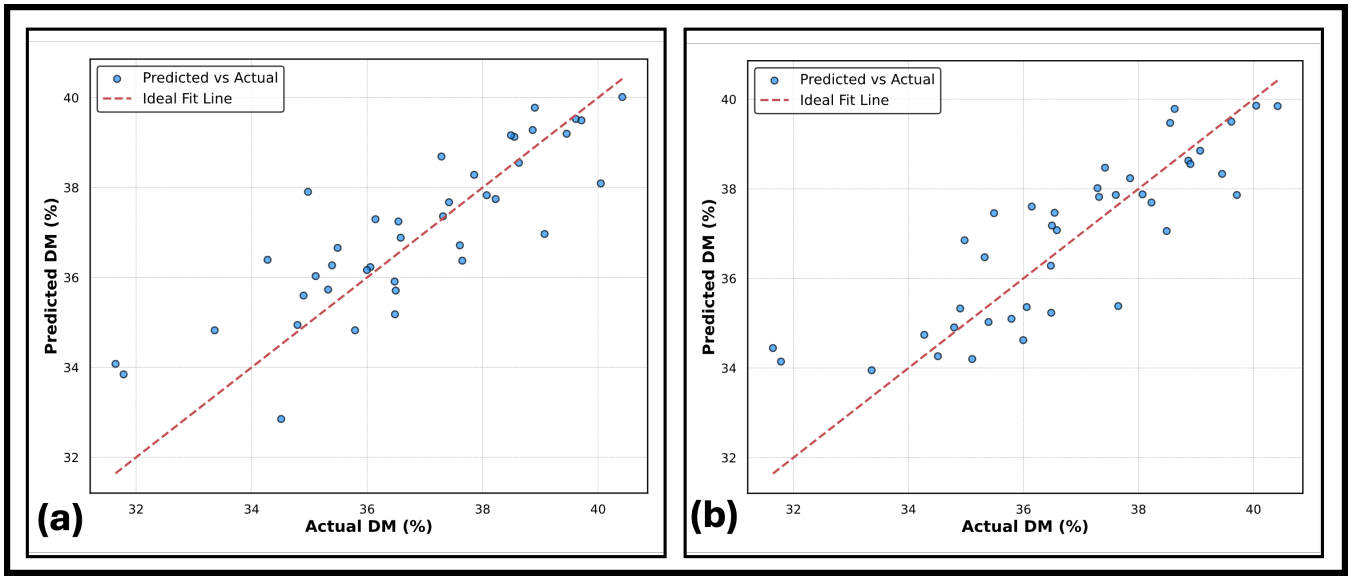

Fig. 3. Predicted vs. actual DM content using iterative random forest regression. (a) Model trained with features at 65 MHz and 199 MHz, evaluated with a 30% test split, achieving $R^2 = 0.785$ and $\mathrm{MSE} = 1.09$. (b) Model trained with features at 0.03 MHz, 3.0 MHz, and 73.5 MHz, evaluated with a 30% test split, achieving $R^2 = 0.791$ and $\mathrm{MSE} = 0.97$.

forest regressor was trained using a 70:30 train-test split, and the R^2 score was computed to evaluate the model's predictive accuracy. The results highlighted 65 MHz and 199 MHz as the best-performing combination, with an R^2 score of 0.78 and an MSE of 1.09 as shown in Fig. 3(a). This means that, among the sampled combinations, these two frequencies allowed the random forest model to achieve the highest predictive accuracy for DM (%).

5) Iterative Random Forest modeling for Three-frequency combinations (IRF-3F): Expanding on the two-frequency modeling, this section explores three-frequency combinations to evaluate whether adding a third frequency improves the random forest model's predictive performance for DM (%).

With 401 available frequencies, the total number of possible three-frequency combinations is 10,726,600. For computational efficiency, a random subset of 5,000 combinations was selected for modeling. Each combination was used to train a random forest model on a 70:30 train-test split, and the R^2 score was recorded to assess model performance.

The highest-performing combination identified was 30 kHz, 3.0 MHz, and 73.5 MHz, achieving an R^2 score of 0.79 and MSE of 0.97 which can be seen from the regression plot in Fig. 3(b). This suggests that incorporating a third frequency provided only a marginal improvement over the best two-frequency combination. Further investigation may be required to determine whether additional frequencies may enhance predictive accuracy or if alternative modeling approaches are more effective.

C. Principal Component Analysis

A key approach in both the linear and random forest regression methods is to treat each frequency as an independent variable. This approach is effective for identifying individual frequencies that are strongly associated with DM content. However, it overlooks the potential for correlation and shared variance among frequencies. While this limitation may not severely impact linear models, it can restrict the performance of more complex models such as random forests, which rely on interactions between features to fully capture patterns in the data [31]. To address this, Principal Component Analysis (PCA) was introduced as a dimensionality reduction technique to uncover the latent structure of the frequency data.

PCA transforms the original high-dimensional dataset into a set of uncorrelated variables known as principal components (PCs). Each component is a linear combination of the original frequency features, with associated loadings that represent the contribution of each frequency to that component. The components are ranked by the amount of variance they capture from the dataset, allowing for the most informative directions of variation to be retained while reducing dimensionality [32], [33].

In this study, three principal components—PC1, PC2, and PC3—were extracted. PC1 accounted for 53.44% of the variance in the frequency data, PC2 explained an additional 18.13%, and PC3 accounted for another 9.32% (Fig. 4a). Together, these three components captured over 70% of the total variance.

Conceptually, each principal component captures a distinct axis of variation. For example, PC1 may reflect variation associated with moisture content, while PC2 may capture secondary effects such as those influenced by temperature or measurement noise. As visualised in the PCA scatter plot (Fig. 4b), moving from right (positive PC1) to left (negative PC1) corresponds with an increase in DM (%). This directional trend suggests that PC1 describes variation closely related to DM, implying that the frequencies with strong contributions to PC1 could be informative predictors of DM content.

To better understand the role of individual frequencies, we first looked at those with the largest absolute loadings in PC1. The highest was 105.5 MHz with loading of 0.0678. We then

This article has been accepted for publication in IEEE Transactions on AgriFood Electronics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFE.2025.3624544

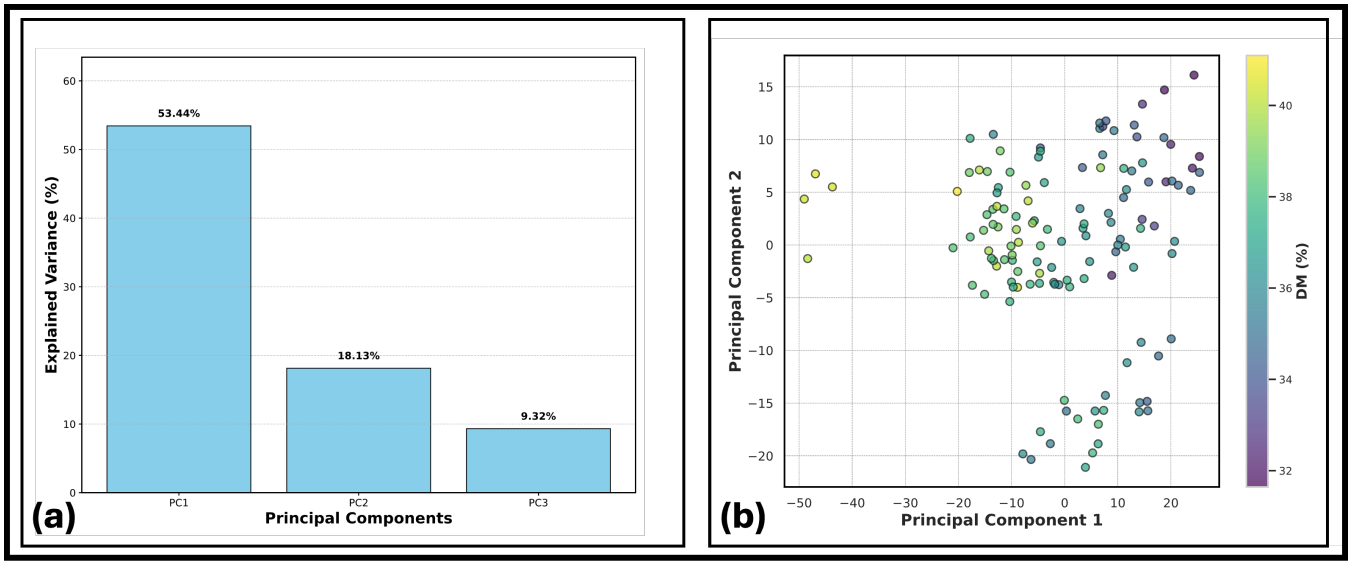

Fig. 4. PCA of RF-R data. (a) Variance explained by the first three principal components, showing that PC1 accounts for 53.44% of the total variance, PC2 for 18.13%, and PC3 for 9.32%. (b) Projection of cassava samples onto the first two principal components PC1 and PC2), color-coded by measured DM content. This visualisation highlights the spread of samples in reduced dimensional space and the relationship between spectral variation and DM.

compared this with frequencies highlighted by other feature selection methods. Their PC1 loadings were very similar: 18.5 MHz (ISLR, 0.0660), 13.5 MHz (IRF, 0.0668), 3.5 MHz (IMLR, 0.0651), 1.0 MHz (IMLR, 0.0600), 73.5 MHz (IMRF, 0.0658), and 65 MHz (IMRF, 0.0637). These loadings were within 12% of the top value at 105.5 MHz, indicating that PCA generally aligns with the other methods in identifying important frequencies. However, PCA is unsupervised and not designed for prediction. For example, although 105.5 MHz had the highest loading, its predictive performance ($R^2 \approx 0.60$) was weaker than that of 18.5 MHz ($R^2 \approx 0.64$). This illustrates an important point: a large PCA loading reflects contribution to overall variance, but it does not necessarily translate into strong predictive power for a specific target. PCA captures variance structure in the data, not features that are optimised for predicting outcomes such as DM (%).

Another approach explored was to combine the frequencies with the highest absolute loadings in PC1 (105.5 MHz) and PC2 (180.5 MHz) as predictor variables in a multiple linear regression model to estimate DM content. The model achieved an R^2 score of 0.61, representing an improvement of less than 2% compared to using 105.5 MHz alone. When a third frequency, corresponding to the highest loading in PC3 (197.5 MHz; Fig. 4a), was added, the three-frequency regression model still yielded an R^2 of 0.61. This indicates that including additional frequencies from PC2 and PC3 provided only marginal, if any, improvement in model performance.

D. Comparison of Computational Efficiency

To compare the various modeling approaches used in this study, execution time serves as the primary metric for computational efficiency. Different ML techniques and frequency selection strategies have been explored to determine their effectiveness in predicting DM%. The methods employed include iterative simple linear regression and iterative random forest models across single-frequency and multi-frequency combinations. Given the computational demands of evaluating all possible frequency combinations, subsampling was applied in certain cases to ensure feasibility.

All computations were performed on two different environments: a local system and a cloud-based platform. The local system is equipped with an Intel® Core™ Ultra 7 165U @ 2.10 GHz processor, 12 cores, and 32 GB RAM, running Windows 11 Pro (Version 24H2, OS Build 26100.3194). In contrast, computations executed in Google Colab were observed to utilise 1 core, 2 logical processors, and 12.67 GB RAM within a Linux-based environment. These differences in computational resources must be considered when interpreting execution times. The execution times recorded for each approach are summarised in Table I.

These results provide a clear indication of the computational cost associated with each method. As expected, random forest models require significantly longer execution times than simple linear regression due to their complexity and ensemble-based nature. The impact of frequency combination size is also evident, with larger search spaces leading to increased computational demand. Notably, evaluating all possible 2-frequency combinations (ISLR-2F) took considerably longer than evaluating 20,000 sampled 3-frequency combinations (ISLR-3F), illustrating the benefits of strategic sub-sampling in maintaining computational efficiency.

The comparison between iterative simple linear regression and random forest models further highlights a trade-off between computational cost and predictive power. While random forest generally offers better predictive performance, it comes at a substantially higher computational cost, particularly when evaluating multi-frequency combinations.

IV. Discussion

Linear regression emerged as a straightforward method for identifying optimal frequencies for DM prediction, offering a

This article has been accepted for publication in IEEE Transactions on AgriFood Electronics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFE.2025.3624544

TABLE I
EXECUTION TIME AND PERFORMANCE METRICS FOR DIFFERENT MODELING APPROACHES

Method	R^2-Score	MSE	Exec. Time
ISLR-1F	0.64	1.65	1s
IRF-1F	0.68	1.673	62s
IMLR-2F (All)	0.71	1.331	300s
IMLR-3F (20k)	0.72	1.276	67s
IMRF-2F (5k)	0.785	1.087	665s
IMRF-3F (5k)	0.786	0.97	670s

clear and interpretable relationship between frequency and DM percentage. The results were relatively simple, with moderate R^2 values indicating a reasonable level of explanatory power. While linear regression successfully highlighted individual frequencies with predictive value, performance improved when multiple frequencies were combined. This reflects the additive effect of different spectral regions contributing complementary information to DM.

PCA further contributed by reducing dimensionality and summarizing variation across the spectral data. Interestingly, some of the principal components with the highest explained variance corresponded to frequencies already identified through linear regression, suggesting consistency between unsupervised and supervised methods. This convergence reinforces the biological relevance of certain frequency bands in capturing internal quality traits like dry matter and starch content. However, PCA does not directly optimise for predictive performance, and in some cases, frequencies with high loading values did not yield the strongest model performance—underscoring the need for complementary feature selection strategies that consider both variance structure and predictive strength.

In contrast, the random forest model provided comparable R^2 values to linear regression but did not substantially improve accuracy despite its capacity to model non-linear interactions. This may be due to the limited number of features or the constrained variability in the dataset. Importantly, the relatively small sample size (131 samples) increases the risk of overfitting, particularly for ensemble models like random forest, thereby limiting model generalisability. Despite this, the convergence of frequency selection across methods and the stability of performance metrics suggest that the findings are robust within the current data scope.

From a deployment perspective, model runtime is more relevant than raw computational complexity. In agricultural settings, especially in remote or low-resource environments, tools for crop quality assessment need to function efficiently on portable or embedded systems. Models that rely on combining multiple features or that require ensemble methods may achieve marginal gains in predictive performance but at the cost of increased processing time and energy consumption. In contrast, simpler models such as single-frequency linear regression or low-component multivariate models can deliver rapid predictions with acceptable accuracy, making them more practical for in-field use.

Several limitations should be acknowledged. The modest dataset size restricts the capacity to fully explore model complexity and validate findings across broader conditions. The spectral data used may not capture the full range of environmental, varietal, or physiological variation encountered in field conditions. Additionally, the absence of independent validation across seasons or geographies means that caution is warranted when extrapolating these findings beyond the current dataset.

Nevertheless, the study demonstrates the feasibility of using low-frequency impedance spectroscopy for non-destructive, low-cost estimation of DM. The methodological framework, starting from single-frequency evaluation, through multivariate regression and PCA, to ensemble modeling, offers a reproducible and interpretable pathway for trait prediction. While this study focused on cassava, the approach is readily generalisable to other crops where internal quality traits such as moisture, oil, or sugar content are not visually apparent but are critical for market value and postharvest performance.

V. CONCLUSION

This work provides a framework for the application of AI/ML to enhance the usability of RF-R in crop and food quality analysis, highlighting the balance between interpretability, computational efficiency, and predictive performance. With further validation and larger datasets, the models developed here could support real-time, field-ready tools for varietal screening, harvest timing, and quality grading—advancing the integration of machine intelligence into precision agriculture.

The results demonstrate that AI and ML techniques, particularly random forest, offer valuable insights into improving crop quality assessment via RF-R. However, these methods require a deeper feature set to realise their full predictive potential. While PCA and linear regression provide a solid foundation for identifying optimal frequencies, more advanced ML models, such as random forest and deep learning models, can exploit these features more effectively once they are sufficiently refined. Despite the promise shown by ML techniques, the primary advantage of linear regression lies in its simplicity and explainability, allowing for clear identification of relationships between frequencies and DM estimation.

Future work could explore the inclusion of additional frequencies or employ other feature selection techniques to identify which combinations of features might lead to a better model. This analysis, however, provided an initial understanding of how a combination of frequency data can serve as predictors for DM %, showing promise but also highlighting the need for further refinement in the feature selection and model-building process.

VI. ACKNOWLEDGEMENT

This research was funded by NESTLER (oNe hEalth SusTainabiLity partnership between EU-AFRICA for food sEcuRity), an EU HORIZON project (https://cordis.europa.eu/project/id/101060762) and the Royal Academy of Engineering (RAEng) Research Fellowship awarded to T Odedeyi.

REFERENCES

1 M. Grossi and B. Riccò, "Electrical impedance spectroscopy (eis) for biological analysis and food characterization: A review," *Journal of sensors and sensor systems*, vol. 6, no. 2, pp. 303–325, 2017.

2 T. Odedeyi, I. Rabbi, C. Poole, and I. Darwazeh, "Estimation of starch content in cassava based on coefficient of reflection measurement," *Frontiers in Food Science and Technology*, vol. 2, p. 878023, 2022.

3 C. Poole and I. Darwazeh, *Microwave Active Circuit Analysis and Design*. Elsevier Science, 2015.

4 S. O. Nelson, "Agricultural applications of dielectric measurements," *IEEE Transactions on Dielectrics and Electrical Insulation*, vol. 13, no. 4, pp. 688–702, 2006.

5 T. Odedeyi and I. Darwazeh, "New insights on application of return loss measurement for starch content estimation in cassava," in *2023 IEEE AFRICON*. IEEE, 2023, pp. 1–3.

6 F. Wang, C. Wang, and S. Song, "A study of starch content detection and the visualization of fresh-cut potato based on hyperspectral imaging," *RSC advances*, vol. 11, no. 22, pp. 13 636–13 643, 2021.

7 E. G. Nkouaya Mbanjo, J. Hershberger, P. Peteti, A. Agbona, A. Ikpan, K. Ogunpaimo, S. I. Kayondo, R. S. Abioye, K. Nafiu, E. O. Alamu *et al.*, "Predicting starch content in cassava fresh roots using near-infrared spectroscopy," *Frontiers in plant science*, vol. 13, p. 990250, 2022.

8 A. López, S. Arazuri, I. García, J. Mangado, and C. Jarén, "A review of the application of near-infrared spectroscopy for the analysis of potatoes," *Journal of agricultural and food chemistry*, vol. 61, no. 23, pp. 5413–5424, 2013.

9 C. Evangelista, L. Basiricò, and U. Bernabucci, "An overview on the use of near infrared spectroscopy (nirs) on farms for the management of dairy cows," *Agriculture*, vol. 11, no. 4, p. 296, 2021.

10 R. Howeler, N. Lutaladio, and G. Thomas, *Save and grow: cassava. A guide to sustainable production intensification*. Fao, 2013.

11 W. F. Breuninger, K. Piyachomkwan, and K. Sriroth, "Tapioca/cassava starch: production and use," in *Starch*. Elsevier, 2009, pp. 541–568.

12 I. M. Zainuddin, A. Fathoni, E. Sudarmonowati, J. R. Beeching, W. Gruissem, and H. Vanderschuren, "Cassava post-harvest physiological deterioration: From triggers to symptoms," *Postharvest Biology and Technology*, vol. 142, pp. 115–123, 2018.

13 S. Chaiwanichsiri, S. Ohnishi, T. Suzuki, R. Takai, and O. Miyawaki, "Measurement of electrical conductivity, differential scanning calorimetry and viscosity of starch and flour suspensions during gelatinisation process," *Journal of the Science of Food and Agriculture*, vol. 81, no. 15, pp. 1586–1591, 2001.

14 T. Odedeyi, C. Poole, X. Liu, A. Kassem, G. Oyebode, R. Ismail, and I. Darwazeh, "A low-cost instrument for estimating the starch content of cassava roots based on the measurement of rf return loss," in *2020 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, 2020, pp. 1–5.

15 T. Odedeyi, A. Issa, C. Poole, and I. Darwazeh, "High-throughput starch content estimation using rf return loss: Theory, analysis and test instrument design," in *2024 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, 2024, pp. 1–5.

16 T. Odedeyi, "Dual-pin impedance probe for crop quality estimation using the rf return loss method," in *Conference on Agrifood Electronics (CAFE), 2024 Xanthi, Greece*. IEEE, 2024, pp. 1–4.

17 A. Hinterleitner, T. Bartz-Beielstein, R. Schulz, S. Spengler, T. Winter, and C. Leitenmeier, "Enhancing feature selection and interpretability in ai regression tasks through feature attribution," *arXiv preprint arXiv:2409.16787*, 2024.

18 K. H. Zou, K. Tuncali, and S. G. Silverman, "Correlation and simple linear regression," *Radiology*, vol. 227, no. 3, pp. 617–628, 2003.

19 D. Fredkin and I. Mayergoyz, "Resonant behavior of dielectric objects (electrostatic resonances)," *Physical Review Letters*, vol. 91, no. 25, p. 253902, 2003.

20 C. Liu, C. Wang, J. Chen, Y. Su, L. Qiao, J. Zhou, and Y. Bai, "Ultrasensitive frequency shifting of dielectric mie resonance near metallic substrate," *Research*, 2022.

21 C. Zhang and Y. Ma, *Ensemble machine learning*. Springer, 2012, vol. 144.

22 M. Fratello, R. Tagliaferri *et al.*, "Decision trees and random forests," *Encyclopedia of bioinformatics and computational biology: ABC of bioinformatics*, vol. 1, no. S3, p. 374, 2018.

23 M. Sivakumar, S. Parthasarathy, and T. Padmapriya, "Trade-off between training and testing ratio in machine learning for medical image processing," *PeerJ Computer Science*, vol. 10, p. e2245, 2024.

24 S. Raschka, "Model evaluation, model selection, and algorithm selection in machine learning," *arXiv preprint arXiv:1811.12808*, 2018.

25 L. Breiman, "Random forest, vol. 45," *Mach Learn*, vol. 1, 2001.

26 S. Han, B. D. Williamson, and Y. Fong, "Improving random forest predictions in small datasets from two-phase sampling designs," *BMC medical informatics and decision making*, vol. 21, no. 1, p. 322, 2021.

27 D. E. Farrar and R. R. Glauber, "Multicollinearity in regression analysis: the problem revisited," *The Review of Economic and Statistics*, pp. 92–107, 1967.

28 N. Shrestha, "Detecting multicollinearity in regression analysis," *American journal of applied mathematics and statistics*, vol. 8, no. 2, pp. 39–42, 2020.

29 J. H. Kim, "Multicollinearity and misleading statistical results," *Korean journal of anesthesiology*, vol. 72, no. 6, pp. 558–569, 2019.

30 Python Software Foundation, *itertools — Functions creating iterators for efficient looping*, Python Software Foundation, 2024, accessed July 2025. [Online]. Available: https://docs.python.org/3/library/itertools.html

31 F. Kherif and A. Latypova, "Principal component analysis," in *Machine learning*. Elsevier, 2020, pp. 209–225.

32 N. Salem and S. Hussein, "Data dimensional reduction and principal components analysis," *Procedia Computer Science*, vol. 163, pp. 292–299, 2019.

33 F. L. Gewers, G. R. Ferreira, H. F. D. Arruda, F. N. Silva, C. H. Comin, D. R. Amancio, and L. d. F. Costa, "Principal component analysis: A natural approach to data exploration," *ACM Computing Surveys (CSUR)*, vol. 54, no. 4, pp. 1–34, 2021.