Journal of Applied and Advanced Research 2024, 9: 1-14
Doi: 10.21839/jaar.2024.v9.8993

Research Article

A random forest-based analysis of cassava mosaic disease-related factors affecting the on-farm livelihoods of cassava farmers

Dickmi Vaillam Claudette*, Tchouamo Isaac Roger

Faculty of Agronomy and Agricultural Sciences, University of Dschang, P.O. Box 96, Dschang, West Region, Cameroon

(Received: April 15, 2024; Revised: June 17, 2024; Accepted: June 19, 2024; Published: June 24, 2024)
*Corresponding Author: Dickmi Vaillam Claudette (E-mail: ngangbaiclaudette@gmail.com)

ABSTRACT

This study aimed to identify key CMD-related factors affecting Cameroon cassava farmers' incomes originating from both the sale of cassava cuttings (V215) and the sale of cassava roots (V216). To achieve this, nine CMD-related variables were used to independently train two Random Forest models. These models were later employed for regression-based prediction of both financial targets V215 and V216. The Random Forest (RF)-based mean absolute percentage error for targets V215 and V216 were 0.19 and 1.25 respectively. The RF-based mean Gaussian deviance for targets V215 and V216 were 0.07 and 0.51 respectively. Based on RF feature importance scores (RFFI), the top 3 factors affecting income originating from the sale of cassava cuttings were found to be: late appearance of symptoms as a difficulty associated with regular field monitoring (RFFI of 0.2594), removal of infected plants as a method of controlling frequent occurrence of viral diseases in respondents' cassava fields (RFFI of 0.1633) and lack of healthy planting material due to frequent occurrence of viral diseases in respondents' cassava fields (RFFI of 0.1495). Also, the top 3 factors affecting income originating from the sale of cassava roots were found to be: the replacement of infected plants with healthy cuttings as a method of controlling the frequent occurrence of viral diseases in respondents' cassava fields (RFFI of 0.1974), decrease in yield due to frequent occurrence of viral diseases in respondents' cassava fields (RFFI of 0.1530) and poor plant growth due to frequent occurrence of viral diseases in respondents' cassava fields (RFFI of 0.1388).

Key words: Cassava mosaic disease, Cassava farmers, Cassava income, Random Forest

INTRODUCTION

Cassava production plays a vital role in the socio-economic landscape of Cameroon, impacting farmers, the local population, and the national economy. For cassava farmers, the crop is often cultivated on a subsistence basis, with most farmers producing for personal consumption rather than commercial purposes (Njukwe *et al.,* 2014). Women are heavily involved in cassava production, often being the primary producers, although they tend to have smaller farm sizes and lower yields compared to men. The majority of cassava farmers are smallholders, with 65% of farmers having less than 1 hectare of land, which limits their economies of scale and income potential (Bilong *et al.,* 2022). Many farmers continue to use traditional methods, such as intercropping and manual processing, which can be labor-intensive and affect yields. For the local population, cassava is a staple food, particularly in the forest regions, providing a significant source of starchy calories and being consumed in various forms (Tize *et al.,* 2021). Cassava is also a good source of carbohydrates, providing energy for the local population. Additionally, it is used in traditional medicine and as a raw material for various products (Evouna *et al.,* 2024). Cassava has cultural and social significance, with many traditional practices and celebrations centered around its production and consumption. In terms of the Cameroon economy, cassava is the second most important crop after rice, in terms of production and consumption. Cassava production provides employment opportunities for many people, particularly women, and contributes to household incomes (Meyo & Liang, 2012). There is also potential for export, particularly to neighboring countries, which could increase foreign exchange earnings and contribute to economic growth (Akiyo, 2013). The government has implemented initiatives to improve cassava production and processing, such as the Nation Program for Development of Roots and Tubers (PNDRT), to enhance the sector's contribution to the economy.

Cassava mosaic disease (CMD) is a significant threat to cassava production in various regions, particularly in Africa (Thuy *et al.,* 2021; Malik *et al.,* 2022; Sheat *et al.,* 2022; Shirima *et al.,* 2022; Chaiyana *et al.,* 2024). The disease is caused by several geminiviruses, including African cassava mosaic virus (ACMV) (Naseem & Winter, 2016; Alabi & Mulenga, 2017), East African cassava mosaic virus (EACMV) (Naseem & Winter, 2016), and East African cassava mosaic Cameroon virus (EACMCMV) (Kouakou *et al.,* 2024). The symptoms of CMD are characterized by leaf curling and distortion, mosaic patterns, yellowing, and stunting, which ultimately lead to reduced plant growth, lowered yields, or even complete loss of the crop. The disease is primarily spread through contaminated cuttings, although whitefly-borne infection can also occur in some regions (Cassava Mosaic Disease, n.d.; Fondong, 2017; Chikoti & Tembo, 2022; Uke *et al.,* 2022;

Hareesh *et al.,* 2023; Sheat & Winter, 2023). The severity of CMD varies across different regions and years. For example, in Burkina Faso, the disease was found to be most prevalent in the Centre-Sud region in 2016, with an incidence of 18.5%, while in 2017, it was highest in the Boucle du Mouhoun region with an incidence of 51.7% (Soro *et al.,* 2021). The severity of the disease also differs, with the lowest severity observed in the East region and the highest in the Sud-Ouest region. Phylogenetic analysis has revealed the presence of three clades of cassava mosaic geminiviruses (CMGs) in Burkina Faso, including ACMV, ACMBFV and EACMCMV (Soro *et al.,* 2021). These viruses are associated with different levels of severity and are often found in co-infections. The management of CMD involves various strategies, including phytosanitation, vector control, breeding for resistance, and genetic engineering (Thresh & Cooter, 2005; Chikoti *et al.,* 2019; Uke *et al.,* 2022). These approaches have been employed in different regions to mitigate the impact of the disease on cassava production. The objective of this work was therefore to identify key CMD-related factors affecting Cameroon cassava farmers' incomes originating from both the sale of cassava cuttings and the sale of cassava roots.

MATERIALS AND METHODS

Dataset

The dataset had six-hundred-and-thirty (630) records originating from respondents residing within four (04) regions of Cameroon – Adamawa, Center, East and South. As summarized in Table 1, there were two target numerical variables – income originating from the sale of cassava cuttings (V215) and income originating from the sale of cassava roots (V216). There were also nine feature binary (yes/no) categorical variables – frequent occurrence of viral diseases in respondents' cassava fields (V341), poor plant growth due to frequent occurrence of viral diseases in respondents' cassava fields (V370), decrease in yield due to frequent occurrence of viral diseases in respondents' cassava fields (V371), lack of healthy planting material due to frequent occurrence of viral diseases in respondents' cassava fields (V372), removal of infected plants as a method of controlling frequent occurrence of viral diseases in respondents' cassava fields (V377), destruction of infected plants as a method of controlling frequent occurrence of viral diseases in respondents' cassava fields (V378), replacement of infected plants with healthy cuttings as a method of controlling frequent occurrence of viral diseases in respondents' cassava fields (V379), use of inputs as a method of controlling frequent occurrence of viral diseases in respondents' cassava fields (V382), and late appearance of symptoms as a difficulty associated with regular field monitoring (V398).

Table 1: Description and codes of variables used in this study

	Code	Variable Description
		TARGETS
1	V215	How much did you raise by selling the cassava cuttings (bag/kg/trucks, etc.)?
2	V216	How much did you raise from the selling of cassava root last cropping season?
		FEATURES
1	V341	What diseases and pests do you frequently see in your cassava fields?/3=Viral (CMD, CBSD)
2	V370	What is the impact of the appearance of these symptoms (CMD) on cassava plants/yield?/1=Poor plant growth
3	V371	What is the impact of the appearance of these symptoms (CMD) on cassava plants/yield?/2=Decrease in yield
4	V372	What is the impact of the appearance of these symptoms (CMD) on cassava plants/yield?/3=Lack of healthy plant material
5	V377	How do you react when your cassava plants show the symptoms shown in photo 2?/1=Removal of infected plants
6	V378	How do you react when your cassava plants show the symptoms shown in photo 2?/2=Destruction of infected plants
7	V379	How do you react when your cassava plants show the symptoms shown in photo 2?/3=Replacement of infected plants with healthy cuttings
8	V382	How do you react when your cassava plants show the symptoms shown in photo 2?/6=Use of inputs
9	V398	Are you experiencing difficulties in monitoring the fields on a regular basis?/4=Late appearance of symptoms

Study sites

In Adamawa, respondents were sourced from Nadeke, Meiganga, Sabongari, Nbaboy and Bounou. In Center, respondents were sourced from Akono, Bikok, Nomayos, Obala, Yemessoa, Mbele, Nkolbibanda, Nkolbibanda 2, Mbankomo, Mendo, Nnom Nnam, Minkama, Ekoadjom, Ngoungoum, Minkotmbem, PK2, Kaya, Nkong Abock, Mvounkeng, Ngoumou, Oveng Centre, Nkoli, Mahomy, Linoyoi, Tamalong, Bomb, Maboye, Boumnyebel Village, Dikonop 1, Makak, Bondjock, Bonde, Lindoi, Dibang, Bogso Village, Mandongwa, Lipombe, Nkolbikok and Koukoum. In East, respondents were sourced from Chefferie, Mbougue, Mekouambe, Ngaikada, STBC, Mbewa, Mission Catholique Djow, Plateau, Derriere Enieg, Derriere Lycée Technique, BLOCK 1, BLOCK 2, Adouma, Peage, Ewankan, Stade, Église, Djenassoume, Yandage 2, *Kaï*gama, Moundi, Meet 2, Bindanang, Sund City, Agofit, Muang 1, Boam, Kpolota, Andom 4, Nguinda, Endom 2, Bamako, Carrefour, Mazabe, Ntankuimb, Doumpi, Megnengue, Pesage, Gare Routiere, Soka, Koak, Obul, Nkoam, Mandjou, Bertoua 2, Sambi, Ndong Mbome, Bogomene, Mboule, Madouma, Aboundoum, Koa and Kak. Finally, in South, respondents were sourced from Mefoup, Alouam, Nyazanga, Midi-me-over, and Mekalat.

Data analysis

The following were used at various stages of the data analysis workflow.

Feature responses

This involved computing the number of positive and negative responses per binary categorical feature.

Chi-square-based feature associations

The Chi-square test, often denoted as X^2, is a powerful statistical tool used to assess associations between categorical variables. Specifically, it helps determine whether the observed frequency distribution of a categorical variable significantly deviates from the expected distribution (Zhang, 2019; Aslam & Smarandache, 2023). Researchers employ Chi-square tests to test hypotheses related to the distribution of categorical variables, such as assessing whether two variables are associated or independent (Gaboardi *et al.,* 2016; Cardona *et al.,* 2020). The Chi-Square Test of Independence, a type of X^2 which is of interest in this study, assesses whether two categorical variables are related to each other. It investigates whether the observed joint frequencies in a contingency table (cross-tabulation) significantly differ from what we would expect if the variables were independent (Benhamou & Melot, 2018; Zhang, 2024).

Random Forest

The concept of random decision forests was first proposed by Salzberg and Heath in 1993 (Heath *et al.,* 1993), with a method that used a randomized decision tree algorithm to generate multiple different trees and then combine them using majority voting. This idea was further developed by Ho in 1995, who established that forests of trees splitting with oblique hyperplanes can gain accuracy as they grow without suffering from overtraining, as long as the forests are randomly restricted to be sensitive to only selected feature dimensions (Ho, 1995). The proper introduction of random forests was made in a paper by Leo Breiman (Breiman, 2001), which combined several ingredients, including bagging, randomized node optimization, and out-of-bag error estimation, to form the basis of the modern practice of random forests.

The Random Forest algorithm operates by constructing a multitude of decision trees at training time. Each tree is grown to the largest extent possible without pruning, and the predictions of the individual trees are combined to determine the final output of the algorithm. The main steps involved in the random forest algorithm are (1) Data Preprocessing: The training data is split into subsets, and a random subset of features is selected for each tree, (2) Building Decision Trees: Decision trees are built for each subset of data, with each tree learning different patterns from the data, (3) Combining Predictions: The predictions from each decision tree are combined using a voting mechanism, such as majority voting, to determine the final output (Utkin & Konstantinov, 2022; Watson *et al.,* 2023Barreñada *et al.,* 2024; Broutin *et al.,* 2024; Curth *et al.,* 2024; Ferry *et al.,* 2024; Surve & Pradhan, 2024; Waltz, 2024).

Random Forest has several advantages that make it a popular choice for many machine learning tasks: (1) Robustness to Overfitting: By combining multiple decision trees, Random Forest reduces the risk of overfitting and improves the model's generalizability, (2) Handling Complex Problems: Random Forest can handle complex problems by combining the predictions of multiple decision trees, (3) Feature Selection: Random Forest can be used for feature selection by analyzing the importance of each feature in the model's predictions and (4) Interpretability: Random Forest provides interpretability through visualization and feature importance analysis, which helps understand the model's decision-making process (Popuri, 2022; Chi *et al.,* 2023; Nam & Han, 2023).

Random Forest is widely used in various machine learning applications, including classification, regression, and feature selection. It is particularly effective in handling high-dimensional data and large-scale datasets, making it a popular choice for many real-world applications. The Random Forest model used in this study is visualized in Figure 1.

Random Forest model evaluation

The Random Forest model used in this study was evaluated, for each of the target variables, using ten (10) metrics – coefficient of determination (R^2) (Piepho, 2018; Jones, 2019; Hawinkel *et al.,* 2024), the mean squared error (MSE) (Das *et al.,* 2004; Kato & Hotta, 2021; Kim *et al.,* 2021; Jin & Montúfar, 2023), the root mean squared error (RMSE) (Zollanvari & Dougherty, 2013; Busch *et al.,* 2014; Huang *et al.,* 2017; Belliardo & Giovannetti, 2020; Zhu, 2022; Reiter & Werner, 2024), the mean absolute error (MAE) (De Myttenaere *et al.,* 2015a, 2016; Qi *et al.,* 2020a, b; Baumgärtner *et al.,* 2023; Wang *et al.,* 2023; Xie, 2024), the mean absolute percentage error (MAPE) (De Myttenaere *et al.,* 2015b), the

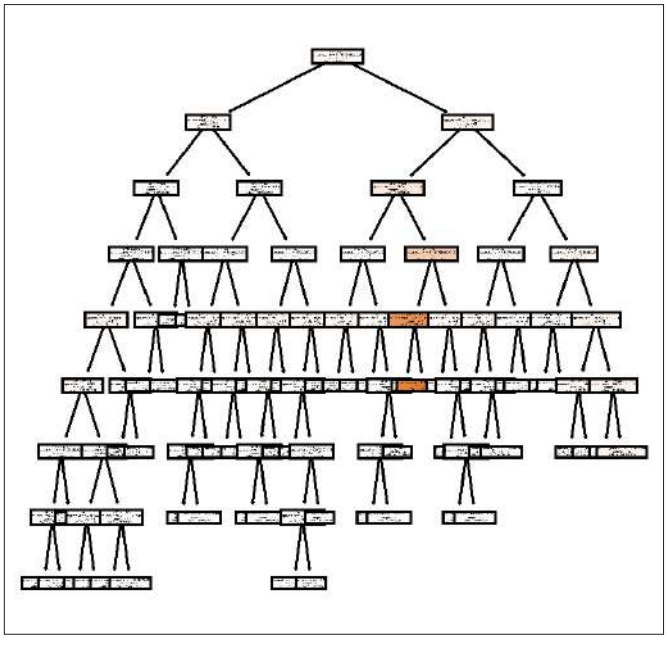

Figure 1: Visualization of the Random Forest model used in this study

maximum error (ME) (Lingasubramanian *et al.,* 2011), the mean pinball loss (MPL) (Sluijterman *et al.,* 2024), the mean gamma deviance (MGD) (Cheema *et al.,* 2023), the mean Poisson deviance (MPD) (Oliveira *et al.,* 2023) and the mean Tweedie deviance (MTD) (Wüthrich & Merz, 2023).

Random Forest partial dependence plots

Random forest partial dependence plots (PDPs) are a powerful tool for understanding how individual features contribute to the predictions made by a random forest model. These plots visualize the marginal effect of a feature on the predicted outcomes of the model, providing valuable insights into how the model is making its predictions. These plots typically show the relationship between a specific feature and the predicted outcome of the model, with the feature on the x-axis and the predicted outcome on the y-axis (Inouye *et al.,* 2020; Moosbauer *et al.,* 2021).

The PDP can be particularly useful for understanding how the model is making predictions, especially when dealing with complex models like random forests. By examining the plot, how the model's predictions change as the value of a particular feature changes can be observed. For example, one might see that as the age of a patient increases, the model becomes more likely to predict that the patient will be readmitted to the hospital. One important consideration when interpreting PDPs is that they are visual descriptions of the model itself, rather than the real-world situation it is trying to model. This means that if the model is not performing well, the PDP will still show how the feature contributed to the model's predictions, but it may not accurately reflect the real-world relationship between the variables. In addition to understanding individual feature effects, PDPs can also be used to explore feature interactions. By plotting the partial dependence of the model on multiple features, one can see how the model's predictions change as the values of multiple features change. This can be particularly useful for identifying complex relationships between features that may not be immediately apparent from other model metrics (Baniecki *et al.,* 2021; Molnar *et al.,* 2021; Xin *et al.,* 2024).

Random Forest residual plots

Random forest residual plots are a crucial tool in model exploration and diagnostics. They help identify different types of issues with model fit or prediction, such as problems with distributional assumptions or the assumed structure of the model. These plots are particularly useful for detecting groups of observations for which a model's predictions are biased and require inspection. One key aspect of random forest residual plots is that they can indicate heteroscedasticity, which is a departure from the assumption of constant variance in residuals. This is less of a concern in random forest models compared to linear regression models because the models reduce the variability of residuals by introducing a bias towards the average (Li *et al.,* 2023). However, it is still important for developers to decide whether this bias is a desirable trade-off for the reduced residual variability. Random forest residual plots can also help in detecting outliers and identifying potential issues with the model. These plots can be used to visualize the relationship between residuals and predicted values, allowing for a more detailed understanding of the model's performance. In addition to identifying issues with the model, random forest residual plots can also be used to monitor the performance of the model over time or across different subsets of the data. This is particularly useful in regression problems where the goal is to predict continuous outcomes. By analyzing the residuals, model developers can refine their models and improve their predictive accuracy (Raymaekers & Rousseeuw, 2021; Warton, 2022).

Random Forest feature importances

The feature importance in Random Forest is typically calculated based on the decrease in impurity or the gain in the information that each feature contributes to the decision-making process. This is done by looking at how much each feature reduces the impurity of the tree nodes across all trees in the forest. The feature importance is then scaled so that the sum of all importance is equal to one, providing a relative measure of the importance of each feature. Random Forest feature importance can be computed using two main methods: mean decrease in impurity (MDI) and permutation feature importance. The MDI method calculates the feature's importance based on the decrease in impurity or the gain in the information that each feature contributes to the decision-making process. This method is fast and efficient but can be biased towards high-cardinality features (Li *et al.,* 2019; Scornet, 2020; Agarwal *et al.,* 2023). The permutation feature importance method, on the other hand, is more robust and less biased. It works by randomly permuting the values of each feature and measuring the decrease in model performance. This method provides a more accurate measure of feature importance but can be computationally expensive for large datasets (Hassan *et al.,* 2021; Chamma *et al.,* 2023; Fumagalli *et al.,* 2023). Random Forest features are useful in several ways. They can help in identifying the most important features in the dataset, which can be used to reduce the dimensionality of the data and improve the model's performance. They can also be used to identify features that are not contributing significantly to the prediction process and can be removed to prevent overfitting.

RESULTS AND DISCUSSION

Feature responses

Without region-based resolution (Figure 2a), 70.47% of respondents noted frequent occurrences of viral diseases in their cassava fields, while 29.53% did not. 69.29% of respondents observed poor plant growth due to frequent occurrences of viral diseases in their cassava fields, while 30.71% did not. 55.31% of respondents noted a decrease in yield due to frequent occurrence of viral diseases in their

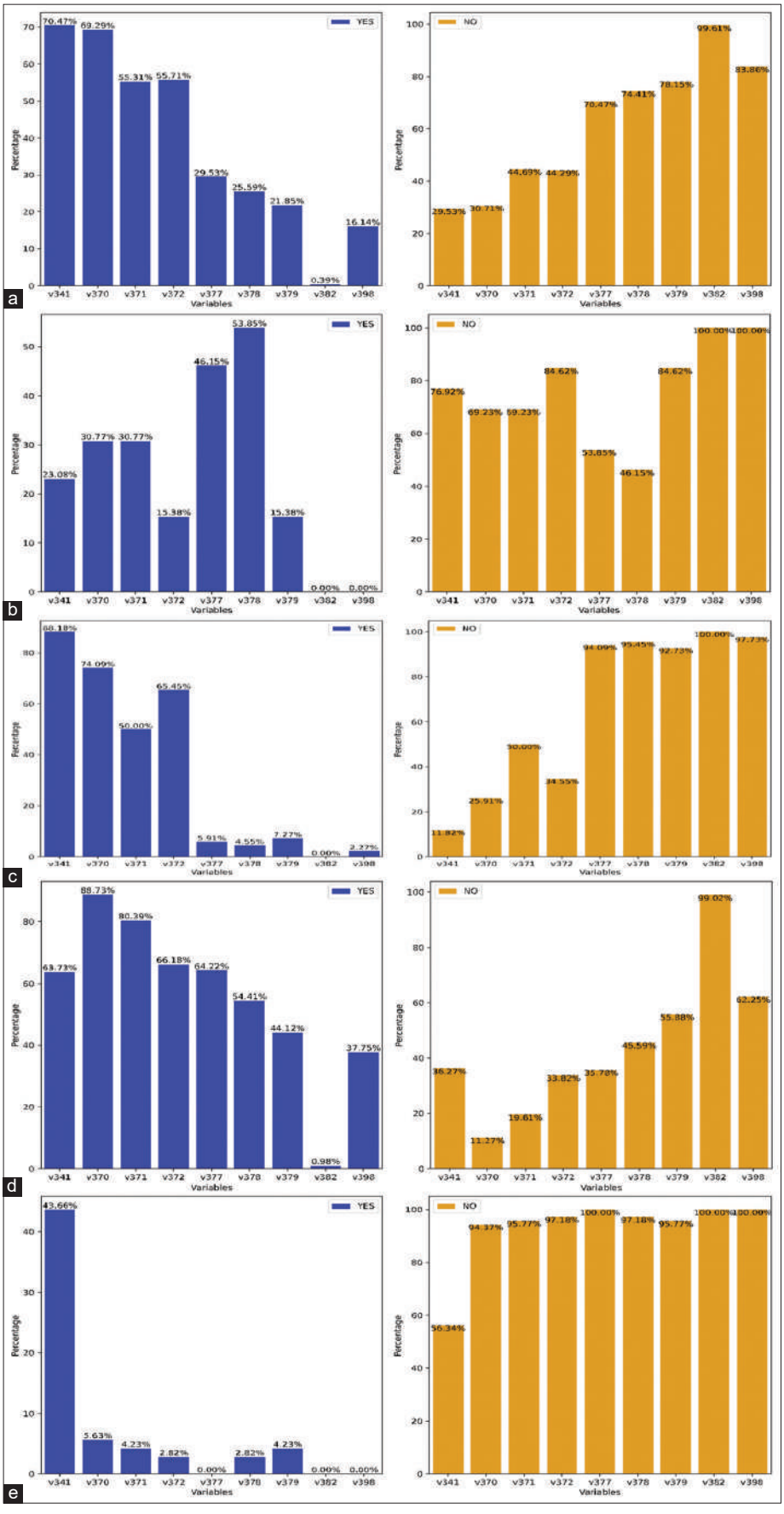

Figure 2: Feature responses. a) Aggregate, b) Adamawa Region, c) Center Region, d) East Region and e) South Region

cassava fields, while 44.69% did not. 55.71% of respondents complained of a lack of healthy planting material due to the frequent occurrence of viral diseases in their cassava fields, while 44.29% did not. 29.53% of respondents reported having practiced the removal of infected plants as a method of controlling the frequent occurrence of viral diseases in their cassava fields, while 70.47% did not. 25.59% of respondents reported having practiced the destruction of infected plants as a method of controlling the frequent occurrence of viral diseases in their cassava fields, while 74.41% did not. 21.85% of respondents reported having practiced the replacement of infected plants with healthy cuttings as a method of controlling the frequent occurrence of viral diseases in their cassava fields, while 78.15% did not. 0.39% of respondents reported having practiced the use of inputs as a method of controlling the frequent occurrence of viral diseases in their cassava fields, while 99.61% did not. Finally, 16.14% of respondents reported late appearance of symptoms as a difficulty associated with regular field monitoring, while 83.86% did not.

In the Adamawa region (Figure 2b), 23.08% of respondents noted frequent occurrences of viral diseases in their cassava fields, while 76.92% did not. 30.77% of respondents observed poor plant growth due to frequent occurrences of viral diseases in their cassava fields, while 69.23% did not. 30.77% of respondents noted a decrease in yield due to the frequent occurrence of viral diseases in their cassava fields, while 69.23% did not. 15.38% of respondents complained of a lack of healthy planting material due to the frequent occurrence of viral diseases in their cassava fields, while 84.62% did not. 46.15% of respondents reported having practiced the removal of infected plants as a method of controlling the frequent occurrence of viral diseases in their cassava fields, while 53.85% did not. 53.85% of respondents reported having practiced the destruction of infected plants as a method of controlling the frequent occurrence of viral diseases in their cassava fields, while 46.15% did not. 15.38% of respondents reported having practiced the replacement of infected plants with healthy cuttings as a method of controlling the frequent occurrence of viral diseases in their cassava fields, while 84.62% did not. None of the respondents reported having practiced the use of inputs as a method of controlling the frequent occurrence of viral diseases in their cassava fields, while 100.00% did not. Finally, none of the respondents reported late appearance of symptoms as a difficulty associated with regular field monitoring, while 100.00% did not.

In the Center region (Figure 2c), 88.18% of respondents noted frequent occurrences of viral diseases in their cassava fields, while 11.82% did not. 74.09% of respondents observed poor plant growth due to the frequent occurrence of viral diseases in their cassava fields, while 25.91% did not. 50.00% of respondents noted a decrease in yield due to the frequent occurrence of viral diseases in their cassava fields, while 50.00% did not. 65.45% of respondents complained of a lack of healthy planting material due to the frequent occurrence of viral diseases in their cassava fields, while 34.55% did not. 5.91% of respondents reported having practiced the removal of infected plants as a method of controlling the frequent occurrence of viral diseases in their cassava fields, while 94.09% did not. 4.55% of respondents reported having practiced the destruction of infected plants as a method of controlling the frequent occurrence of viral diseases in their cassava fields, while 95.45% did not. 7.27% of respondents reported having the practiced replacement of infected plants with healthy cuttings as a method of controlling the frequent occurrence of viral diseases in their cassava fields, while 92.73% did not. 0.00% of respondents reported having practiced the use of inputs as a method of controlling the frequent occurrence of viral diseases in their cassava fields, while 100.00% did not. Finally, 2.27% of respondents reported late appearance of symptoms as a difficulty associated with regular field monitoring, while 97.73% did not.

In the East region (Figure 2d), 63.73% of respondents noted frequent occurrence of viral diseases in their cassava fields, while 36.27% did not. 88.73% of respondents observed poor plant growth due to frequent occurrences of viral diseases in their cassava fields, while 11.27% did not. 80.39% of respondents noted a decrease in yield due to frequent occurrence of viral diseases in their cassava fields, while 19.61% did not. 66.18% of respondents complained of a lack of healthy planting material due to the frequent occurrence of viral diseases in their cassava fields, while 33.82% did not. 64.22% of respondents reported having practiced the removal of infected plants as a method of controlling the frequent occurrence of viral diseases in their cassava fields, while 35.78% did not. 54.41% of respondents reported having practiced the destruction of infected plants as a method of controlling the occurrence of viral diseases in their cassava fields, while 45.59% did not. 44.12% of respondents reported having practiced the replacement of infected plants with healthy cuttings as a method of controlling the frequent occurrence of viral diseases in their cassava fields, while 55.88% did not. 0.98% of respondents reported having practiced the use of inputs as a method of controlling the frequent occurrence of viral diseases in their cassava fields, while 99.02% did not. Finally, 37.75% of respondents reported late appearance of symptoms as a difficulty associated with regular field monitoring, while 62.25% did not.

In the South region (Figure 2e), 43.66% of respondents noted frequent occurrence of viral diseases in their cassava fields, while 56.34% did not. 5.63% of respondents observed poor plant growth due to frequent occurrence of viral diseases in their cassava fields, while 94.37% did not. 4.23% of respondents noted decrease in yield due to frequent occurrence of viral diseases in their cassava fields, while 95.77% did not. 2.82% of respondents complained of lack of healthy planting material due to frequent occurrence of viral diseases in their cassava fields, while 97.18% did not. 0.00% of respondents reported having practiced removal of infected plants as a method of controlling frequent occurrence of viral diseases in their cassava fields, while 100.00% did not. 2.82% of respondents reported having practiced destruction of infected plants as a method of controlling frequent occurrence of viral

diseases in their cassava fields, while 97.18% did not. 4.23% of respondents reported having practiced replacement of infected plants with healthy cuttings as a method of controlling frequent occurrence of viral diseases in their cassava fields, while 95.77% did not. 0.00% of respondents reported having practiced the use of inputs as a method of controlling frequent occurrence of viral diseases in their cassava fields, while 100.00% did not. Finally, 0.00% of respondents reported late appearance of symptoms as a difficulty associated with regular field monitoring, while 100.00% did not.

Chi-square-based feature associations (alpha = 0.05)

Without region-based resolution (Figure 3a), there were significant associations between frequent occurrence of viral diseases in respondents' cassava fields and poor plant growth due to frequent occurrence of viral diseases in respondents' cassava fields (7.74E-07), frequent occurrence of viral diseases in respondents' cassava fields and lack of healthy planting material due to frequent occurrence of viral diseases in respondents' cassava fields (1.9E-04), frequent occurrence of viral diseases in respondents' cassava fields and removal of infected plants as a method of controlling frequent occurrence of viral diseases in respondents' cassava fields (9.29E-06), frequent occurrence of viral diseases in respondents' cassava fields and destruction of infected plants as a method of controlling frequent occurrence of viral diseases in respondents' cassava fields (6.70E-05), and frequent occurrence of viral diseases in respondents' cassava fields and replacement of infected plants with healthy cuttings as a method of controlling frequent occurrence of viral diseases in respondents' cassava fields (3.2E-04). Poor plant growth due to frequent occurrence of viral diseases in respondents' cassava fields was also significantly associated with the following: decrease in yield due to frequent occurrence of viral diseases in respondents' cassava fields (4.82E-36), lack of healthy planting material due to frequent occurrence of viral diseases in respondents' cassava fields (1.77E-43), removal of infected plants as a method of controlling the frequent occurrence of viral diseases in respondents' cassava fields (2.13E-08), destruction of infected plants as a method of controlling the frequent occurrence of viral diseases in respondents' cassava fields (6.76E-06), replacement of infected plants with healthy cuttings as a method of controlling the frequent occurrence of viral diseases in respondents' cassava fields (5.14E-06), and late appearance of symptoms as a difficulty associated with regular field monitoring (1.30E-05).

Decrease in yield due to frequent occurrence of viral diseases in respondents' cassava fields was also significantly associated with the following: lack of healthy planting material due to the frequent occurrence of viral diseases in respondents' cassava fields (5.41E-23), removal of infected plants as a method of controlling the frequent occurrence of viral diseases in respondents' cassava fields (2.34E-09), destruction of infected plants as a method of controlling frequent occurrence of viral diseases in respondents' cassava fields (1.50E-12), replacement of infected plants with healthy cuttings as a method of controlling frequent occurrence of viral diseases in respondents' cassava fields (4.14E-12), and late appearance of symptoms as a difficulty associated with regular field monitoring (7.84E-08). Lack of healthy planting material due to frequent occurrence of viral diseases in respondents' cassava fields was also significantly associated with the following: destruction of infected plants as a method of controlling frequent occurrence of viral diseases in respondents' cassava fields (0.023348042), replacement of infected plants with healthy cuttings as a method of controlling frequent occurrence of viral diseases in respondents' cassava fields (5.48E-05), and late appearance of symptoms as a difficulty associated with regular field monitoring (0.001857773).

Removal of infected plants as a method of controlling frequent occurrence of viral diseases in respondents' cassava fields was also significantly associated with the following: destruction of infected plants as a method of controlling frequent occurrence of viral diseases in respondents' cassava fields (3.27E-66), replacement of infected plants with healthy cuttings as a method of controlling frequent occurrence of viral diseases in respondents' cassava fields (2.43E-46), and late appearance of symptoms as a difficulty associated with regular field monitoring (7.54E-14). Destruction of infected plants as a method of controlling frequent occurrence of viral diseases in respondents' cassava fields was also significantly associated with the following: replacement of infected plants with healthy cuttings as a method of controlling frequent occurrence of viral diseases in respondents' cassava fields (2.03E-40), and late appearance of symptoms as a difficulty associated with regular field monitoring (3.36E-17). Finally, the replacement of infected plants with healthy cuttings as a method of controlling the frequent occurrence of viral diseases in respondents' cassava fields was also significantly associated with the late appearance of symptoms as a difficulty associated with regular field monitoring (7.29E-13).

In the Adamawa region (Figure 3b), poor plant growth due to the frequent occurrence of viral diseases in respondents' cassava fields was significantly associated with the removal of infected plants as a method of controlling frequent occurrence of viral diseases in respondents' cassava fields (0.046197685), decrease in yield due to frequent occurrence of viral diseases in respondents' cassava fields was significantly associated with removal of infected plants as a method of controlling the frequent occurrence of viral diseases in respondents' cassava fields (0.046197685), and removal of infected plants as a method of controlling the frequent occurrence of viral diseases in respondents' cassava fields was significantly associated with destruction of infected plants as a method of controlling frequent occurrence of viral diseases in respondents' cassava fields (0.011326189).

In the Center region (Figure 3c), frequent occurrence of viral diseases in respondents' cassava fields was significantly associated with lack of healthy planting material due to frequent occurrence of viral diseases in respondents' cassava

Figure 3: Chi-square-based feature associations. a) Aggregate, b) Adamawa Region, c) Center Region, d) East Region and e) South Region

fields (0.015369056), poor plant growth due to frequent occurrence of viral diseases in respondents' cassava fields was significantly associated with both decrease in yield due to frequent occurrence of viral diseases in respondents' cassava fields (1.28E-11) and lack of healthy planting material due to frequent occurrence of viral diseases in respondents' cassava fields (2.48E-26); decrease in yield due to frequent occurrence of viral diseases in respondents' cassava fields was significantly associated with lack of healthy planting material due to frequent occurrence of viral diseases in respondents' cassava fields (1.11E-05), destruction of infected plants as a method of controlling frequent occurrence of viral diseases in respondents' cassava fields (0.00357946) and replacement of infected plants with healthy cuttings as a method of controlling frequent occurrence of viral diseases in respondents' cassava fields (0.004292751); removal of infected plants as a method of controlling frequent occurrence of viral diseases in respondents' cassava fields was significantly associated with both destruction of infected plants as a method of controlling frequent occurrence of viral diseases in respondents' cassava fields (2.45E-21) and replacement of infected plants with healthy cuttings as a method of controlling frequent occurrence of viral diseases in respondents' cassava fields (5.32E-13). Finally, the destruction of infected plants as a method of controlling the frequent occurrence of viral diseases in respondents' cassava fields was significantly associated with the replacement of infected plants with healthy cuttings as a method of controlling the frequent occurrence of viral diseases in respondents' cassava fields (2.70E-09).

In East region (Figure 3d), frequent occurrence of viral diseases in respondents' cassava fields was significantly associated with the following: decrease in yield due to frequent occurrence of viral diseases in respondents' cassava fields (0.003305109), removal of infected plants as a method of controlling frequent occurrence of viral diseases in respondents' cassava fields (1.71E-17), destruction of infected plants as a method of controlling frequent occurrence of viral diseases in respondents' cassava fields (4.74E-16) and replacement of infected plants with healthy cuttings as a method of controlling frequent occurrence of viral diseases in respondents' cassava fields (5.57E-10); poor plant growth due to frequent occurrence of viral diseases in respondents' cassava fields was significantly associated with decrease in yield due to frequent occurrence of viral diseases in respondents' cassava fields (9.72E-05) and lack of healthy planting material due to frequent occurrence of viral diseases in respondents' cassava fields (0.007436266). Decrease in yield due to frequent occurrence of viral diseases in respondents' cassava fields was significantly associated with lack of healthy planting material due to frequent occurrence of viral diseases in respondents' cassava fields (1.92E-07) and removal of infected plants as a method of controlling frequent occurrence of viral diseases in respondents' cassava fields (0.076588499). Lack of healthy planting material due to frequent occurrence of viral diseases in respondents' cassava fields was significantly associated with replacement of infected plants with healthy cuttings as a method of controlling frequent occurrence of viral diseases in respondents' cassava fields (0.003047493). Removal of infected plants as a method of controlling frequent occurrence of viral diseases in respondents' cassava fields was significantly associated with destruction of infected plants as a method of controlling frequent occurrence of viral diseases in respondents' cassava fields (7.84E-19) and replacement of infected plants with healthy cuttings as a method of controlling frequent occurrence of viral diseases in respondents' cassava fields (3.98E-14). Destruction of infected plants as a method of controlling frequent occurrence of viral diseases in respondents' cassava fields was significantly associated with replacement of infected plants with healthy cuttings as a method of controlling frequent occurrence of viral diseases in respondents' cassava fields (3.79E-12) and late appearance of symptoms as a difficulty associated with regular field monitoring (0.005355511). Finally, replacement of infected plants with healthy cuttings as a method of controlling frequent occurrence of viral diseases in respondents' cassava fields was significantly associated with late appearance of symptoms as a difficulty associated with regular field monitoring (0.013097939).

In the South region (Figure 3e), only poor plant growth due to frequent occurrence of viral diseases in respondents' cassava fields was significantly associated with destruction of infected plants as a method of controlling frequent occurrence of viral diseases in respondents' cassava fields (1.59E-05).

Random forest model evaluation

Table 2 summarizes the model evaluation metrics for targets V215 and V216 – income originating from the sale of cassava cuttings and income originating from the sale of cassava roots, respectively.

Worthy of note is that the MAPE (with the main advantage being statistically valid comparability across datasets and scales) for targets V215 and V216 were 0.19 and 1.25 respectively, and the MGD (with the main advantage being the best description for the spread in a Gaussian probability distribution) for targets V215 and V216 were 0.07 and 0.51 respectively.

Random forest partial dependence plots

As presented in Figure 4, the lack of healthy planting material due to the frequent occurrence of viral diseases in respondents' cassava fields, removal of infected plants as a method of controlling the frequent occurrence of viral diseases in respondents' cassava fields and destruction of infected plants as a method of controlling frequent occurrence of viral diseases in respondents' cassava fields show a direct relationship with both income originating from the sale of cassava cuttings and income originating from the sale of cassava roots, indicating that as these predictors increase, the target variables tend to increase as well. A decrease in yield due to the frequent occurrence of viral diseases in respondents' cassava fields and the use of inputs as a method of controlling the frequent occurrence of viral diseases in respondents' cassava fields suggest a negligible influence on the target variables, implying that changes in these predictors do not significantly affect the predictions for income originating from the sale of cassava cuttings and income originating from the sale of cassava roots. Removal of infected plants as a method of controlling the frequent occurrence of viral diseases in respondents' cassava fields and destruction of infected plants as a method of controlling the frequent occurrence of viral diseases in respondents' cassava fields demonstrate strong positive trends, indicating a substantial impact on both incomes originating from the sale of cassava cuttings and income originating from the sale of cassava roots as these predictors increase. Noteworthy is

Table 2: Model evaluation metrics for targets V215 and V216

Metric	V215	V216
R^2	-22.8737691	-0.297488173
MSE	35346426748	1.84165E+11
MAE	47076.34437	275624.9925
RMSE	188006.454	429144.667
MAPE	0.19625629	1.257583697
ME	1358849.531	1513037.292
MPL	23538.17219	137812.4962
MGD	0.074820798	0.516049077
MPD	43011.81952	259659.9998
MTD	35346426748	1.84165E+11

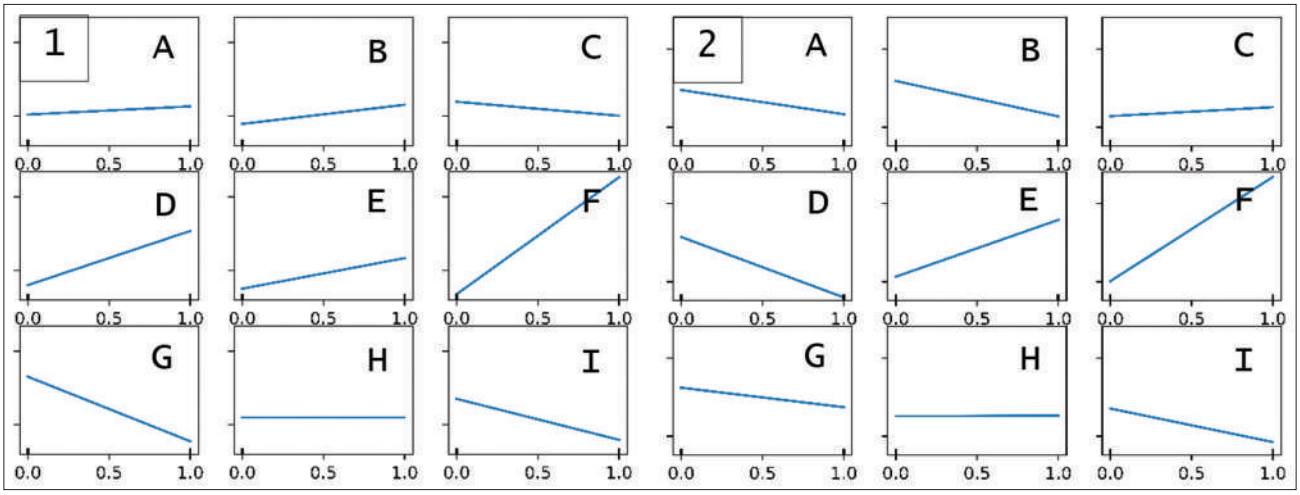

Figure 4: Partial dependence plots for targets 1) V215 and 2) V216. A) V341, B) V370, C) V371, D) V372, E) V377, F) V378, G) V379, H) V382 and I) V398

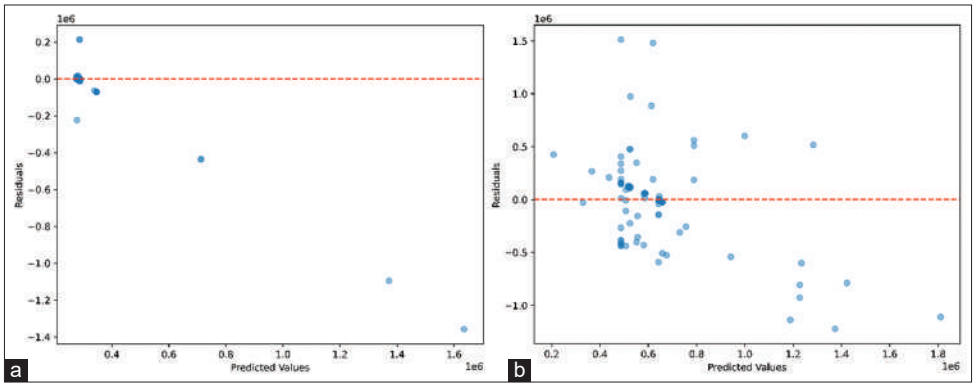

Figure 5: Residual plots for targets a) V215 and b) V216

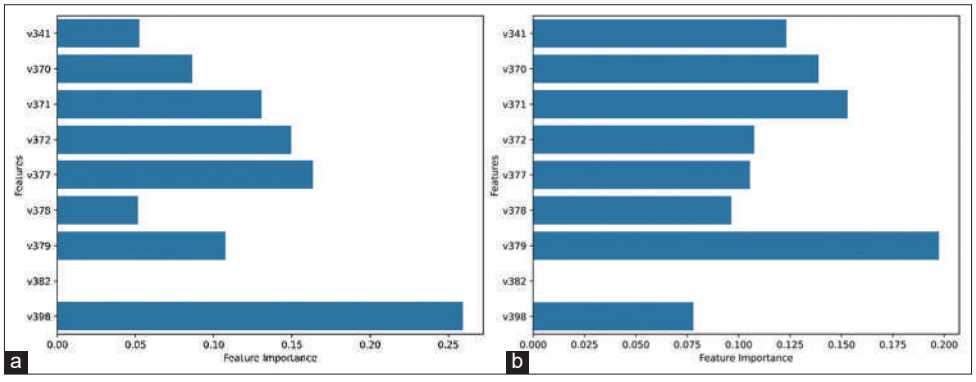

Figure 6: Random forest feature importance scores for targets a) V215 and b) V216

the fact that strong drivers of the predicted outcomes for both income originating from the sale of cassava cuttings and income originating from the sale of cassava roots could imply that strategies focusing on these predictors might be effective in influencing the target variables. Conversely, the flat trends indicate that these predictors do not contribute significantly to the variation in income originating from the sale of cassava cuttings and income originating from the sale of cassava roots. This could mean that resources allocated to these variables might not yield substantial changes in the predicted outcomes.

Random forest residual plots

The residual plots for variables V215 and V216 (Figure 5) provide a visual representation of the discrepancies between the predicted and actual values obtained from a predictive model. In the context of Random Forest regression analysis, these plots are crucial for diagnosing the model's

Table 3: Random forest feature importance scores for targets V215 and V216

Features	V215	V216
V341	0.0523	0.1232
V370	0.0860	0.1388
V371	0.1304	0.1530
V372	0.1495	0.1075
V377	0.1633	0.1055
V378	0.0516	0.0964
V379	0.1073	0.1974
V382	0.0001	0.0002
V398	0.2594	0.0779

performance and identifying any systematic errors that may exist.

For V215 (Figure 5a): the points are widely scattered, indicating a significant variance in the residuals. This suggests that the model's predictions for V215 are not consistently close to the actual values. The lack of a discernible pattern or trend in the residuals could imply that the model is not capturing all the necessary information or that there may be outliers influencing the predictions. The wide dispersion could also be indicative of non-linearity in the relationship between the predictors and the target variable, V215, or heteroscedasticity, where the variance of the residuals is not constant across the range of predicted values. For V216 (Figure 5b): The residuals are more tightly clustered around the zero line, which is a positive sign that the model's predictions for V216 are generally more accurate. Despite the tighter clustering, there is still some variability present, which means there are still prediction errors that need to be addressed. Compared to V215, the model appears to perform better for V216, as evidenced by the closer grouping of residuals around the zero line.

While both models show room for improvement, the model for V216 seems to be more reliable. Further analysis may be required to fine-tune the models, possibly by exploring additional predictors, applying transformation techniques, or considering different modeling approaches to better capture the underlying patterns in the data.

Random forest feature importances

Table 3 and Figure 6 numerically and visually (respectively) summarize Random Forest feature importance scores for targets V215 and V216 – income originating from the sale of cassava cuttings and income originating from the sale of cassava roots, respectively.

The top 3 factors affecting income originating from the sale of cassava cuttings include the late appearance of symptoms as a difficulty associated with regular field monitoring (0.2594), removal of infected plants as a method of controlling frequent occurrence of viral diseases in respondents' cassava fields (0.1633) and lack of healthy planting material due to frequent occurrence of viral diseases in respondents' cassava fields (0.1495).

The top 3 factors affecting income originating from the sale of cassava roots include replacement of infected plants with healthy cuttings as a method of controlling frequent occurrence of viral diseases in respondents' cassava fields (0.1974), decrease in yield due to frequent occurrence of viral diseases in respondents' cassava fields (0.1530) and poor plant growth due to frequent occurrence of viral diseases in respondents' cassava fields (0.1388).

Collecting the data used for this study in rural areas of Cameroon presented several challenges that impacted the quality and effectiveness of the data collection efforts. Described here are the major limitations associated with data collection during this study, (1) Infrastructure and Connectivity: inadequate infrastructure and limited access to electricity and internet connectivity hindered data collection since tablets were the data collection tools. The lack of reliable power sources affected electronic data collection. (2) Human Resource Capacity: the authors registered a lack of trained personnel for data management and utilization. (3) Logistical Challenges: rural areas were characterized by dispersed populations, making it difficult to reach all locations for data collection. (4) Data Quality Assurance: ensuring data accuracy and consistency was challenging due to limited supervision and resources.

CONCLUSION

This study has shown that the continuous reliance of cassava farmers on vegetative propagation-based cassava planting material is principally responsible for the bulk of the financial losses which they experience, and so an adoption of seed-based cassava planting material would go a long way to secure their on-farm livelihood in cassava production.

REFERENCES

Agarwal, A., Kenney, A. M., Tan, Y. S., Tang, T. M., & Yu, B. (2023). *MDI+: A Flexible Random Forest-Based Feature Importance Framework*. arXiv:2307.01932. https://arxiv.org/abs/2307.01932v1

Akiyo, S. (2013). Cassava Processing and Marketing by Rural Women in the Central Region of Cameroon. *African Study Monographs, 34*(4), 203-219. https://doi.org/10.14989/185092

Alabi, O. J., & Mulenga, R. M. (2017). African cassava mosaic virus (African cassava mosaic). *CABI Compendium*, 2535. https://doi.org/10.1079/cabicompendium.2535

Aslam, M., & Smarandache, F. (2023). Chi-square test for imprecise data in consistency table. *Frontiers in Applied Mathematics and Statistics, 9*, 1279638. https://doi.org/10.3389/fams.2023.1279638

Baniecki, H., Kretowicz, W., & Biecek, P. (2023). Fooling Partial Dependence via Data Poisoning. In M. R. Amini, S. Canu, A. Fischer, T. Guns, P. K. Novak, & G. Tsoumakas (Eds.), *Machine Learning and Knowledge Discovery in Databases* (Vol. 13715) Cham, Switzerland: Springer. https://doi.org/10.1007/978-3-031-26409-2_8

Barreñada, L., Dhiman, P., Timmerman, D., Boulesteix, A.-L., & Van Calster, B. (2024). *Understanding random forests and overfitting:*

A visualization and simulation study. arXiv:2402.18612. https://doi.org/10.48550/arXiv.2402.18612

Baumgärtner, L., Herzog, R., Schmidt, S., & Weiß, M. (2023). *The Proximal Map of the Weighted Mean Absolute Error*. arXiv:2209.13545. https://doi.org/10.48550/arXiv.2209.13545

Belliardo, F., & Giovannetti, V. (2020). Achieving Heisenberg scaling with maximally entangled states: An analytic upper bound for the attainable root mean square error. *Physical Review A, 102*(4), 042613. https://doi.org/10.1103/PhysRevA.102.042613

Benhamou, E., & Melot, V. (2018). *Seven proofs of the Pearson Chi-squared independence test and its graphical interpretation*. arXiv:1808.09171. https://doi.org/10.48550/arXiv.1808.09171

Bilong, E. G., Abossolo-Angue, M., Ajebesone, F. N., Anaba, B. D., Madong, B. À., Nomo, L. B., & Bilong, P. (2022). Improving soil physical properties and cassava productivity through organic manures management in the southern Cameroon. *Heliyon, 8*(6), e09570. https://doi.org/10.1016/j.heliyon.2022.e09570

Breiman, L. (2001). Random Forests. *Machine Learning, 45*(1), 5-32. https://doi.org/10.1023/A:1010933404324

Broutin, N., Devroye, L., Lugosi, G., & Oliveira, R. I. (2024). *Subtractive random forests* arXiv:2210.10544. https://doi.org/10.48550/arXiv.2210.10544

Busch, P., Lahti, P., & Werner, R. F. (2014). Quantum root-mean-square error and measurement uncertainty relations. *Reviews of Modern Physics, 86*(4), 1261-1281. https://doi.org/10.1103/RevModPhys.86.1261

Cardona, L. A. S., Vargas-Cardona, H. D., Navarro González, P., Cardenas Peña, D. A., & Orozco Gutiérrez, Á. Á. (2020). Classification of Categorical Data Based on the Chi-Square Dissimilarity and t-SNE. *Computation, 8*(4), 4. https://doi.org/10.3390/computation8040104

Cassava Mosaic Disease. (n.d.). Cassava Mosaic Disease: A Curse to Food Security in Sub-Saharan Africa. Retrieved from https://www.apsnet.org/edcenter/apsnetfeatures/Pages/cassava.aspx

Chaiyana, A., Khiripet, N., Ninsawat, S., Siriwan, W., Shanmugam, M. S., & Virdis, S. G. P. (2024). Mapping and predicting cassava mosaic disease outbreaks using earth observation and meteorological data-driven approaches. *Remote Sensing Applications: Society and Environment, 35*, 101231. https://doi.org/10.1016/j.rsase.2024.101231

Chamma, A., Thirion, B., & Engemann, D. A. (2023, December 18). *Variable Importance in High-Dimensional Settings Requires Grouping*. arXiv:2312.10858. https://arxiv.org/abs/2312.10858v1

Cheema, M., Amin, M., Mahmood, T., Faisal, M., Brahim, K., & Elhassanein, A. (2023). Deviance and Pearson Residuals-Based Control Charts with Different Link Functions for Monitoring Logistic Regression Profiles: An Application to COVID-19 Data. *Mathematics, 11*(5), 5. https://doi.org/10.3390/math11051113

Chi, C.-M., Fan, Y., & Lv, J. (2023). *FACT: High-Dimensional Random Forests Inference*. arXiv:2207.01678). https://doi.org/10.48550/arXiv.2207.01678

Chikoti, P. C., & Tembo, M. (2022). Expansion and impact of cassava brown streak and cassava mosaic diseases in Africa: A review. *Frontiers in Sustainable Food Systems, 6*, 1076364. https://doi.org/10.3389/fsufs.2022.1076364

Chikoti, P. C., Mulenga, R. M., Tembo, M., & Sseruwagi, P. (2019). Cassava mosaic disease: A review of a threat to cassava production in Zambia. *Journal of Plant Pathology, 101*(3), 467-477. https://doi.org/10.1007/s42161-019-00255-0

Curth, A., Jeffares, A., & van der Schaar, M. (2024). *Why do Random Forests Work? Understanding Tree Ensembles as Self-Regularizing Adaptive Smoothers*. arXiv:2402.01502. https://doi.org/10.48550/arXiv.2402.01502

Das, K., Jiang, J., & Rao, J. N. K. (2004). Mean squared error of empirical predictor. *The Annals of Statistics, 32*(2), 818-840. https://doi.org/10.1214/009053604000000201

De Myttenaere, A., Golden, B., Grand, B. L., & Rossi, F. (2015a). *Using the Mean Absolute Percentage Error for Regression Models*. arXiv:1506.04176. https://doi.org/10.48550/arXiv.1506.04176

De Myttenaere, A., Golden, B., Grand, B. L., & Rossi, F. (2016). Mean Absolute Percentage Error for regression models. *Neurocomputing, 192*, 38-48. https://doi.org/10.1016/j.neucom.2015.12.114

De Myttenaere, A., Grand, B. L., & Rossi, F. (2015b). *Empirical risk minimization is consistent with the mean absolute percentage error*. arXiv:1509.02357. https://doi.org/10.48550/arXiv.1509.02357

Evouna, J. S. M., Molua, E. L., Choumbou, R. F. D., & Kambiet, P. L. K. (2024). Structure and performance of cassava markets: Challenges of food security and connecting small farmers to markets in Cameroon. *Frontiers in Sustainable Food Systems, 8*, 1353565. https://doi.org/10.3389/fsufs.2024.1353565

Ferry, J., Fukasawa, R., Pascal, T., & Vidal, T. (2024). *Trained Random Forests Completely Reveal your Dataset*. arXiv:2402.19232. https://doi.org/10.48550/arXiv.2402.19232

Fondong, V. N. (2017). The Search for Resistance to Cassava Mosaic Geminiviruses: How Much We Have Accomplished, and What Lies Ahead. *Frontiers in Plant Science, 8*, 408. https://doi.org/10.3389/fpls.2017.00408

Fumagalli, F., Muschalik, M., Hüllermeier, E., & Hammer, B. (2023). Incremental Permutation Feature Importance (iPFI): Towards Online Explanations on Data Streams. *Machine Learning, 112*, 4863-4903. https://doi.org/10.1007/s10994-023-06385-y

Gaboardi, M., woo Lim, H., Rogers, R., & Vadhan, S. (2016). *Differentially Private Chi-Squared Hypothesis Testing: Goodness of Fit and Independence Testing*. arXiv:1602.03090. https://doi.org/10.48550/arXiv.1602.03090

Hareesh, P. S., Resmi, T. R., Sheela, M. N., & Makeshkumar, T. (2023). Cassava mosaic disease in South and Southeast Asia: Current status and prospects. *Frontiers in Sustainable Food Systems, 7*, 1086660. https://doi.org/10.3389/fsufs.2023.1086660

Hassan, A., Paik, J. H., Khare, S., & Hassan, S. A. (2021). *PPFS: Predictive Permutation Feature Selection*. arXiv:2110.10713. https://arxiv.org/abs/2110.10713v1

Hawinkel, S., Waegeman, W., & Maere, S. (2024). The out-of-sample R^2: Estimation and inference. *The American Statistician, 78*(1), 15-25. https://doi.org/10.1080/00031305.2023.2216252

Heath, D. G., Kasif, S., & Salzberg, S. (1993). Induction of Oblique Decision Trees. *International Joint Conference on Artificial Intelligence*.

Ho, T. K. (1995). Random decision forests. *Proceedings of 3rd International Conference on Document Analysis and Recognition, 1*, 278-282. https://doi.org/10.1109/ICDAR.1995.598994

Huang, M. L., Kerman, R., & Spektor, S. (2017). *An estimate of the root mean square error incurred when approximating an $f \in L^2(R)$ by a partial sum of its Hermite series*. arXiv:1709.03039. https://doi.org/10.48550/arXiv.1709.03039

Inouye, D. I., Leqi, L., Kim, J. S., Aragam, B., & Ravikumar, P. (2020). *Automated Dependence Plots*. arXiv:1912.01108v3. https://arxiv.org/abs/1912.01108v3

Jin, H., & Montúfar, G. (2023). *Implicit Bias of Gradient Descent for Mean Squared Error Regression with Two-Layer Wide Neural Networks*. arXiv:2006.07356. https://doi.org/10.48550/arXiv.2006.07356

Jones, T. (2019). *A Coefficient of Determination for Probabilistic Topic Models*. arXiv:1911.11061. https://doi.org/10.48550/arXiv.1911.11061

Kato, S., & Hotta, K. (2021). *MSE Loss with Outlying Label for Imbalanced Classification*. arXiv:2107.02393. https://doi.org/10.48550/arXiv.2107.02393

Kim, T., Oh, J., Kim, N., Cho, S., & Yun, S.-Y. (2021). *Comparing Kullback-Leibler Divergence and Mean Squared Error Loss in Knowledge Distillation*. arXiv:2105.08919. https://doi.org/10.48550/arXiv.2105.08919

Kouakou, B. S. M., Yoboué, A. A. N., Pita, J. S., Mutuku, J. M., Otron, D. H., Kouassi, N. K., Kouassi, K. M., Vanié-Léabo, L. P. L., Ndougonna, C., Zouzou, M., & Sorho, F. (2024). Gradual Emergence of East African cassava mosaic Cameroon virus in Cassava Farms in Côte d'Ivoire. *Agronomy, 14*(3), 3. https://doi.org/10.3390/agronomy14030418

Li, W., Cook, D., Tanaka, E., & VanderPlas, S. (2023). *A Plot is Worth a Thousand Tests: Assessing Residual Diagnostics with the Lineup Protocol*. arXiv:2308.05964v2. https://arxiv.org/abs/2308.05964v2

Li, X., Wang, Y., Basu, S., Kumbier, K., & Yu, B. (2019). *A Debiased MDI Feature Importance Measure for Random Forests*. arXiv:1906.10845v2. https://arxiv.org/abs/1906.10845v2

Lingasubramanian, K., Alam, S. M., & Bhanja, S. (2011). Maximum Error Modeling for Fault-Tolerant Computation using Maximum a posteriori (MAP) Hypothesis. *Microelectronics Reliability, 51*(2), 485-501. https://doi.org/10.1016/j.microrel.2010.07.156

Malik, A. I., Sophearith, S., Delaquis, E., Cuellar, W. J., Jimenez, J., & Newby, J. C. (2022). Susceptibility of Cassava Varieties to Disease Caused by Sri Lankan Cassava Mosaic Virus and Impacts on Yield by Use of Asymptomatic and Virus-Free Planting Material. *Agronomy, 12*(7), 7. https://doi.org/10.3390/agronomy12071658

Meyo, E. S. M., & Liang, D. (2012). Gap Analysis of Cassava Sector in Cameroon. *International Journal of Economics and Management Engineering, 6*(11), 2792-2799.

Molnar, C., Freiesleben, T., König, G., Casalicchio, G., Wright, M. N., & Bischl, B. (2021). Relating the Partial Dependence Plot and Permutation Feature Importance to the Data Generating Process. In L. Longo (Eds.), *Explainable Artificial Intelligence* (Vol. 1901) Cham, Switzerland: Springer. https://doi.org/10.1007/978-3-031-44064-9_24

Moosbauer, J., Herbinger, J., Casalicchio, G., Lindauer, M., & Bischl, B. (2021, November 8). *Explaining Hyperparameter Optimization via Partial Dependence Plots*. arXiv:2111.04820v2. https://arxiv.org/abs/2111.04820v2

Nam, Y., & Han, S. (2023). *Random Forest Variable Importance-based Selection Algorithm in Class Imbalance Problem*. arXiv:2312.10573. https://doi.org/10.48550/arXiv.2312.10573

Naseem, S., & Winter, S. (2016). Quantification of African cassava mosaic virus (ACMV) and East African cassava mosaic virus (EACMV-UG) in single and mixed infected Cassava (*Manihot esculenta* Crantz) using quantitative PCR. *Journal of Virological Methods, 227*, 23-32. https://doi.org/10.1016/j.jviromet.2015.10.001

Njukwe, E., Onadipe, O., Amadou Thierno, D., Hanna, R., Kirscht, H., Maziya-Dixon, B. B., Araki, S., & Ngue-Bissa, T. (2014). Cassava processing among smallholder farmers in Cameroon: Opportunities and challenges. *International Journal of Agricultural Policy and Research, 2*(4), 113-124.

Oliveira, N. L., Lei, J., & Tibshirani, R. J. (2023). *Unbiased Test Error Estimation in the Poisson Means Problem via Coupled Bootstrap Techniques*. arXiv:2212.01943. https://doi.org/10.48550/arXiv.2212.01943

Piepho, H.-P. (2018). *A Coefficient of Determination (R2) for Linear Mixed Models*. arXiv:1805.01124. https://doi.org/10.48550/arXiv.1805.01124

Popuri, S. K. (2022). *An Approximation Method for Fitted Random Forests*. arXiv:2207.02184. https://doi.org/10.48550/arXiv.2207.02184

Qi, J., Du, J., Siniscalchi, S. M., Ma, X., & Lee, C.-H. (2020a). Analyzing Upper Bounds on Mean Absolute Errors for Deep Neural Network Based Vector-to-Vector Regression. *IEEE Transactions on Signal Processing, 68*, 3411-3422. https://doi.org/10.1109/TSP.2020.2993164

Qi, J., Du, J., Siniscalchi, S. M., Ma, X., & Lee, C.-H. (2020b). On Mean Absolute Error for Deep Neural Network Based Vector-to-Vector Regression. *IEEE Signal Processing Letters, 27*, 1485-1489. https://doi.org/10.1109/LSP.2020.3016837

Raymaekers, J., & Rousseeuw, P. J. (2021). *Silhouettes and quasi residual plots for neural nets and tree-based classifiers*. arXiv:2106.08814v2. https://arxiv.org/abs/2106.08814v2

Reiter, S., & Werner, S. W. R. (2024). *Interpolatory model order reduction of large-scale dynamical systems with root mean squared error measures*. arXiv:2403.08894. https://doi.org/10.48550/arXiv.2403.08894

Scornet, E. (2020). *Trees, forests, and impurity-based variable importance*. arXiv:2001.04295v3. https://arxiv.org/abs/2001.04295v3

Sheat, S., & Winter, S. (2023). Developing broad-spectrum resistance in cassava against viruses causing the cassava mosaic and the cassava brown streak diseases. *Frontiers in Plant Science, 14*, 1042701. https://doi.org/10.3389/fpls.2023.1042701

Sheat, S., Zhang, X., & Winter, S. (2022). High-Throughput Virus Screening in Crosses of South American and African Cassava Germplasm Reveals Broad-Spectrum Resistance against Viruses Causing Cassava Brown Streak Disease and Cassava Mosaic Virus Disease. *Agronomy, 12*(5), 5. https://doi.org/10.3390/agronomy12051055

Shirima, R. R., Wosula, E. N., Hamza, A. A., Mohammed, N. A., Mouigni, H., Nouhou, S., Mchinda, N. M., Ceasar, G., Amour, M., Njukwe, E., & Legg, J. P. (2022). Epidemiological Analysis of Cassava Mosaic and Brown Streak Diseases, and Bemisia tabaci in the Comoros Islands. *Viruses, 14*(10), 10. https://doi.org/10.3390/v14102165

Sluijterman, L., Kreuwel, F., Cator, E., & Heskes, T. (2024). *Composite Quantile Regression With XGBoost Using the Novel Arctan Pinball Loss*. arXiv:2406.02293. https://doi.org/10.48550/arXiv.2406.02293

Soro, M., Tiendrébéogo, F., Pita, J. S., Traoré, E. T., Somé, K., Tibiri, E. B., Néya, J. B., Mutuku, J. M., Simporé, J., & Koné, D. (2021). Epidemiological assessment of cassava mosaic disease in Burkina Faso. *Plant Pathology, 70*(9), 2207-2216. https://doi.org/10.1111/ppa.13459

Surve, T., & Pradhan, R. (2024). *Example-based Explanations for Random Forests using Machine Unlearning*. arXiv:2402.05007. https://doi.org/10.48550/arXiv.2402.05007

Thresh, J. M., & Cooter, R. J. (2005). Strategies for controlling cassava mosaic virus disease in Africa. *Plant Pathology, 54*(5), 587-614. https://doi.org/10.1111/j.1365-3059.2005.01282.x

Thuy, C. T. L., Lopez-Lavalle, L. A. B., Vu, N. A., Hy, N. H., Nhan, P. T., Ceballos, H., Newby, J., Tung, N. B., Hien, N. T., Tuan, L. N., Hung, N., Hanh, N. T., Trang, D. T., Ha, P. T. T., Ham, L. H., Hoi Pham, X., Quynh, D. T. N., Rabbi, I. Y.,

Kulakow, P. A., & Zhang, X. (2021). Identifying New Resistance to Cassava Mosaic Disease and Validating Markers for the CMD2 Locus. *Agriculture, 11*(9), 9. https://doi.org/10.3390/agriculture11090829

Tize, I., Fotso, A. K., Nukenine, E. N., Masso, C., Ngome, F. A., Suh, C., Lendzemo, V. W., Nchoutnji, I., Manga, G., Parkes, E., Kulakow, P., Kouebou, C., Fiaboe, K. K. M., & Hanna, R. (2021). New cassava germplasm for food and nutritional security in Central Africa. *Scientific Reports*, *11*, 7394. https://doi.org/10.1038/s41598-021-86958-w

Uke, A., Tokunaga, H., Utsumi, Y., Vu, N. A., Nhan, P. T., Srean, P., Hy, N. H., Ham, L. H., Lopez-Lavalle, L. A. B., Ishitani, M., Hung, N., Tuan, L. N., Van Hong, N., Huy, N. Q., Hoat, T. X., Takasu, K., Seki, M., & Ugaki, M. (2022). Cassava mosaic disease and its management in Southeast Asia. *Plant Molecular Biology, 109*(3), 301-311. https://doi.org/10.1007/s11103-021-01168-2

Utkin, L. V., & Konstantinov, A. V. (2022). *Attention and Self-Attention in Random Forests*. arXiv:2207.04293. https://doi.org/10.48550/arXiv.2207.04293

Waltz, N. (2024). *Grafting: Making Random Forests Consistent*. arXiv:2403.06015. https://doi.org/10.48550/arXiv.2403.06015

Wang, X., Hua, Y., Kodirov, E., Clifton, D. A., & Robertson, N. M. (2023). *IMAE for Noise-Robust Learning: Mean Absolute Error Does Not Treat Examples Equally and Gradient Magnitude's Variance Matters*. arXiv:1903.12141. https://doi.org/10.48550/arXiv.1903.12141

Warton, D. I. (2022). *Global simulation envelopes for diagnostic plots in regression models*. arXiv:2208.01811v2. https://arxiv.org/abs/2208.01811v2

Watson, D. S., Blesch, K., Kapar, J., & Wright, M. N. (2023). *Adversarial random forests for density estimation and generative modeling*. arXiv:2205.09435. https://doi.org/10.48550/arXiv.2205.09435

Wüthrich, M. V., & Merz, M. (2023). Selected Topics in Deep Learning. In M. V. Wüthrich & M. Merz (Eds.), *Statistical Foundations of Actuarial Learning and its Applications* (pp. 453-535) Cham, Switzerland: Springer. https://doi.org/10.1007/978-3-031-12409-9_11

Xie, P. (2024). *Hyb Error: A Hybrid Metric Combining Absolute and Relative Errors*. arXiv:2403.07492. https://doi.org/10.48550/arXiv.2403.07492

Xin, X., Hooker, G., & Huang, F. (2024). *Why You Should Not Trust Interpretations in Machine Learning: Adversarial Attacks on Partial Dependence Plots*. arXiv:2404.18702v2. https://arxiv.org/abs/2404.18702v2

Zhang, Q. (2019). A Class of Association Measures for Categorical Variables Based on Weighted Minkowski Distance. *Entropy, 21*(10), 10. https://doi.org/10.3390/e21100990

Zhang, Q. (2024). *On the properties of distance covariance for categorical data: Robustness, sure screening, and approximate null distributions*. arXiv:2403.17882. https://doi.org/10.48550/arXiv.2403.17882

Zhu, W. (2022). *Statistical parameters for assessing environmental model performance related to sample size: Case study in ocean color remote sensing*. arXiv:2208.05743. https://doi.org/10.48550/arXiv.2208.05743

Zollanvari, A., & Dougherty, E. R. (2013). *Moments and Root-Mean-Square Error of the Bayesian MMSE Estimator of Classification Error in the Gaussian Model*. arXiv:1310.1519. https://doi.org/10.48550/arXiv.1310.1519