Cocoa beans moisture content prediction using Machine Learning Model based on the color image features

Joel E AKO[1,2]
Joel.Ako-Ekissi@insa-rennes.fr

Camille E. N'ZI[1]
camille.nzi@inphb.ci

Kidiyo KPALMA[2]
Kidiyo.Kpalma@insa-rennes.fr

(1)Institut National Polytechnique Félix Houphouët Boigny (INP-HB), Unité Mixte de Recherche et d'innovation (UMRI) en Science des Technologies d'Ingenieur (STI), BP 1093 Yamoussoukro, Côte d'Ivoire ;
(2)Univ Rennes, CNRS, Institut National des Sciences Appliquées (INSA), IETR (Institut d'Electronique et des Technologies du numéRique) - UMR 6164, F-35000 Rennes, France.

ABSTRACT

The moisture content of cocoa beans is an essential factor in their quality. Modeling it during drying is still problematic due to the wide variation in drying conditions and the wide variation in cocoa bean varieties. This article aims to investigate the possibility of modeling the moisture content of cocoa beans as a function of RGB images features of unshelled cocoa beans. The approach is to extract features, analyze them and then use the most relevant ones to study Machine Learning models. Features are extracted by calculating mean, standard deviation, energy, entropy, kurtosis and skewness of the components of the rgb (RGB normalized), HSV, L*a*b*, YCbCr color spaces without the brightness components. These features are extracted from 4 types of samples, namely 10, 30, 50 and 70 bean samples per image. Features analysis using the F-test and RReliefF methods shows that the features based on the energy and entropy of the components rg, yb, Cr, Cb, a*, b* and h* are fairly relevant for predicting the water content of cocoa beans. However, they are highly correlated. The selected predictors allow the analysis of linear models, such as Ridge Regression (RR), PLS Regression (PLSR) and non-linear models, such as polynomial, Support Vector Regression (SVR) with rbf kernel, and Decision Trees Regression (DTR). Except RR and PLSR, the other models were preceded by a principal component analysis (PCA) to handle the collinearity problem. The non-linear models give good predictions for the training dataset, with coefficients of determination R^2 ranging from 0.94 to 0.96 and RMSE from 3.85 to 4.81. However, there is a significant difference between these results and the predictions of the new datasets. RR and PLSR are stable models, but their predictions are less than non-linear ones. It is therefore possible to predict the moisture content of cocoa beans from the features of RGB images.

Keywords

cocoa beans, Moisture content, color features, F-test, RReliefF, Regression, Machine Learning

1 INTRODUCTION

The moisture content of a product is the amount of water present in this product. It is important for the microbiological and nutritional properties of food products, as well as for regulatory and economic aspects. As a result, determining moisture content is one of the most frequent analyses carried out in the food industry. If products are to be stored for long periods, they need to be dried to a certain water content. In the case of cocoa, after harvesting the ripe cocoa pods, fresh cacao beans are fermented and dried immediately after fermentation to safe moisture content from around 60% to 7-8% (ISO 2451/2014 standard) [DJE09] to facilitate storage, transport and guarantee the quality of the beans. Too high a moisture content can encourage the development of mold and alter the quality of the final product, while too low a moisture content can make the beans brittle and difficult to process. It can cause damage that contributes to the depreciation of bean quality [HUM10]. Predicting moisture content during drying is therefore an optimum solution for ensuring quality drying. The moisture content prediction during drying requires a non-destructive solution. However, the implementation of such a solution is very complex due to the instability of drying conditions and the diversity of cocoa beans. Despite this, researchers propose a few solutions. These include modeling drying kinetics [DJE09], [HII11], [IGO15], [KAR18], [CAS23], predicting moisture content by Near-Infrared spectral

[HAS18] and designing artificial dryers [KAV21]. Despite this research, modeling bean moisture content during drying is still problematic, because the models developed depend on drying conditions. This work involves exploring image features to design a model that predict the moisture content of cocoa beans during the drying process. The aim is to identify image features that are not affected by drying conditions and to design a model for this prediction. Our approach consists of two main steps. The first step involves selecting the discriminant predictors and the optimal quantity of beans by using two variable selection methods: the F-test and the Regression Relief Features selection (RReliefF) algorithm. The second step is devoted to studying linear and non-linear Machine Learning models, such as Ridge Regression (RR), PLS Regression (PLSR), polynomial, Support Vector Regression (SVR) with RBF kernel and Decision Trees Regression (DTR), the selected relevant features.

The remainder of this work is organized as follows: Section 2 presents the related solutions. Section 3 and 4 describe the proposed approach and experiments and the datasets, respectively. Section 5 presents the results and discussion and section 6 the main conclusions.

2 RELATED WORKS

The related works concern predicting or modeling the moisture content during drying. The models already developed are based on drying kinetics and Near-Infrared spectral.

Drying kinetics-based modeling

Drying kinetic modeling solutions are based on mathematical and artificial neural network models. Hii C. L. et al. have used Fick's theoretical model to study the drying kinetics of cocoa beans. They obtain coefficient of determination R^2 of training data ranging from 0.9845 to 0.9976. But they mention that the drying process is highly unsteady due to the fluctuating ambient conditions [HII09]. A. Djedjro et al. evaluate a suitable drying mathematical model for describing the drying curves. Among the mathematical models studied, the logarithmic model satisfactorily described the drying behavior of cocoa beans with a coefficient of determination 0.976 and RMSE 0.0128 [DJE09]. Nogbou A. et al. described the behavior of cocoa beans in predicting their moisture content during intermittent microwave drying at different power levels (450 W, 600 W, 700 W). They proposed a recurrent artificial neural network model using drying time, microwave power and moisture content as inputs. They obtained a coefficient of determination ranging between 0.9967 and 0.9993. [IGO15]. Daouda K. et al. proposed a mathematical model of the evolution of cocoa beans moisture content as a function of time using an artificial neural network during the sun drying. They found the multilayer perceptron with two neurons on the input layer to be the most suitable. The coefficient of determination of the linear regression between observed and predicted water content values was 0.99 [KAR18]. Eduardo Castillo et al. fitted a diffusion approximation model using nonlinear regression to the moisture ratio of the CCN51 cocoa bean with the drying time for the constant drying temperatures of $40, 50, 60$, and $70°C$. The coefficient of determination for all cases was 0.9999 with RMSE 0.0044 [CAS23].

Near-infrared spectrum-based modeling

Hashimoto et al built PLS regression models from near-infrared diffuse reflectance spectrum for the prediction of several cocoa bean quality parameters including water content. The coefficient of determination of moisture content prediction is 0.67. [HAS18].

The majority of papers found focuses on Drying kinetics. The advantage of these solutions is that they give a good prediction of moisture content for the training data. The disadvantage is that they depend on drying conditions, i.e. temperature and drying time. However, under natural drying conditions, time and temperature are highly unsteady due to the fluctuating ambient conditions [HII09]. The other solution, which is independent of theses conditions, uses near-infrared spectra, which has a low prediction rate. The proposed approach uses image features to propose model independent of drying conditions.

3 PROPOSED SOLUTION

The proposed solution is based on prediction of cocoa beans moisture content using color features. It involves acquiring images of batches of cocoa beans during drying at regular time intervals to designing the best Machine Learning model for predicting moisture content. The differents stages of the proposed solution are shown in the block diagram in figure 1.

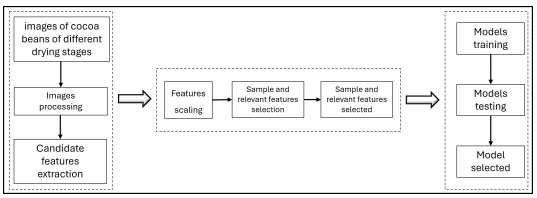


Figure 1: Proposed solution steps

3.1 Image processing

The image processing involves extracting cacao beans from the blue acquisition background. Then, the RGB images are segmented using color thresholding in L*a*b* space. A region of interest (ROI) are created firstly with Matlab Color Thresholder app [MATSG24], then apply morphological opening and closing with the optimal structuring element(disk) to perfect the edges of the extracted beans (figure 2).

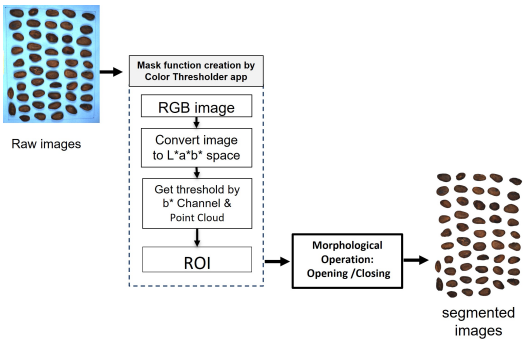


Figure 2: Image segmentation

3.2 Extraction of color-based features

Features are extracted from *RGB* images, using statistical Moments such as mean, standard deviation, energy, entropy, kurtosis and skewness. These methods were applied to the color components of the *rgb* [RAS06], $HSV, YCrCb$ and $L^*a^*b^*$ spaces, and the chromatic components in spherical coordinates θ and ϕ [RAS06], without the luminance components. We also have the components *rg* (red-green) and *yb* (yellow-blue) derived from *rgb* space [WAN14] and C^* and h^* derived from $L^*a^*b^*$ space. $HSV, YCbCr$ and $L^*a^*b^*$ space components are derived from *RGB* images, using the corresponding MATLAB functions. Expressions for the other components (from equation (1) to equation (9) and the statistical Moments (from equation (10) to equation (15)) are following, where R, G, B are the components of *RGB* space, N is the number of bean pixels in the image, A_i is the gray level of pixel i and h_A is the normalised histogram of gray level A of the image I.

$$\mathrm{r} = \frac{R}{R+G+B} \tag{1}$$

$$\mathrm{g} = \frac{G}{R+G+B} \tag{2}$$

$$\mathrm{b} = \frac{B}{R+G+B} \tag{3}$$

$$\mathrm{rg} = r - g \tag{4}$$

$$\mathrm{yb} = \frac{r}{2} + \frac{g}{2} - b \tag{5}$$

$$C^* = \sqrt{a^{*2} + b^{*2}} \tag{6}$$

$$h^* = \arctan\left(\frac{a^*}{b^*}\right) \tag{7}$$

$$\theta = \arctan\left(\frac{G}{R}\right) \tag{8}$$

$$\phi = \arcsin\left(\frac{\sqrt{R^2+G^2}}{\sqrt{R^2+G^2+B^2}}\right) \tag{9}$$

$$\mathrm{Mean} = \frac{1}{N}\sum_{i=1}^{N} A_i \tag{10}$$

$$\mathrm{SD} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(A_i - Mean)^2} \tag{11}$$

$$\mathrm{Energy} = \sum_{A=0}^{255} h_A^2 \tag{12}$$

$$\mathrm{Entropy} = -\sum_{A=0}^{255} h_A ln(h_A) \tag{13}$$

$$\mathrm{Kurtosis} = \frac{1}{SD^4}\sum_{A=0}^{255}(A - Mean)^4 h_A \tag{14}$$

$$\mathrm{Skewness} = \frac{1}{SD^3}\sum_{A=0}^{255}(A - Mean)^3 h_A \tag{15}$$

3.3 Variable selection methods

Variable selection is a step that precedes model design. It allows to analyze the potential explanatory and selection variables most relevant to a model design. It also provides the necessary information on each of the explanatory variables for better use and interpretation in a model. To select discriminating predictors, we use two different methods. The F-test method to assess the significance of candidate variables and Regression Relief Features selection (RReliefF) algorithm to assess their relevance.

F-test

The F-test is a statistical test that compares the variances of two samples, or the ratio of variances between several samples. It is often used to test equality of means in an analysis of variance, or to test the goodness of fit of a regression model. The F-test is based on the F-statistic. It is the ratio of the variance explained by the model to the residual variance. The observed p_{value} are used to interpret the F-test. Higher the F-statistic, the smaller the p_{value}, thus better the model fits data. We use it to assess the goodness of fit of each candidate variable in a linear regression model with moisture content. This involves examining the importance of each candidate color feature, then ranking them using the p_{value} of the F-test statistics. The score for each candidate variable is determined by the following relationship [MATFT24], [OME14]:

$$\mathrm{S} = -\log(p_{value}) \tag{16}$$

Regression Relief Features selection (RReliefF) algorithm

The RReliefF algorithm is an extension of the Relief algorithm, which is a variable selection method based on assigning weights to variables. It detects relevant variables by considering interactions between variables and noise in the data. It also penalizes the predictors

that give different values to neighbors with the same response values, and rewards predictors that give different values to neighbors with different response values. However, it uses intermediate weights to compute the final predictor weights. Then, it calculates the predictor weights W_j after fully updating all the intermediate weights [ROB97].

$$\mathrm{Wj} = \frac{W_{(dy \wedge dj)}}{W_{dy}} - \frac{W_{dj} - W_{(dy \wedge dj)}}{m - W_{dy}} \tag{17}$$

W_{dy} and W_{dj} are the weights of having different values for the response y and predictor x_j, respectively. $W_{(dy \wedge dj)}$ is the weight of having different response y and different values for the response y values and different values for the predictor x_j. m is the number of iterations.

The selected variables will be used to analyse the prediction of water content using Machine Learning models.

3.4 Machine Learning regression models

Different types of regression models are explored, namely: Ordinary Least Squares, Kernel Support Vector, Decision Trees, Ridge and Partial Least Squares models.

Ordinary Least Squares regression (OLSR)

We analyse the Multiple linear Regression (MLR) and Polynomial Regression (PR) Models. The sample regression model has the form [MONT21]:

$$y_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \varepsilon_i \tag{18}$$

Where x_{ij} and y_i are the k predictors and the response respectively of ith observation. The parameters β_j, $j = 0, 1, ..., k$ are the regression coefficients. OLS is the most popular estimation method; its purpose is to find the unbiased coefficients $\beta = (\beta_0, \beta_1, \ldots, \beta_p)^\top$ which minimize the residual sum of squares:

$$RSS(\beta) = \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 \tag{19}$$

Support Vector Regression (SVR) with the radial basis function (RBF) kernel

SVR is based on the Support Vector Machine algorithm. It also based on the computation of a linear regression function in a multiple variables feature space where the input data can be used via a non-linear regression function. Unlike OLSR that aim to minimize the error between the predicted and actual values, SVR aims to fit as many instances as possible within a margin while limiting violations of the margin and controlling the margin width, in other words find a hyperplane that best fits as many data points as possible while minimizing the margin violations. The margin is defined as the region between the hyperplane and the support vectors. A nonlinear function has the form [CAS20]:

$$f(x_i) = \omega^T \Phi(x_i) + b \tag{20}$$

Given training vectors $x_i \in R^p$, $i = 1, .., N$, and a vector $y \in R^N$ SVR solves the following primal problem:

$$min_{(\omega, b, \zeta, \zeta^*)} \left(\frac{1}{2}\omega^T \omega + C \sum_{i=1}^{N} (\zeta_i + \zeta_i^*) \right) \tag{21}$$

subject to $y_i - \omega^T \Phi(x_i) - b \leq \varepsilon + \zeta_i$,

$\omega^T \Phi(x_i) + b - y_i \leq \varepsilon + \zeta_i^*$; $\zeta_i, \zeta_i^* \geq 0, i = 1, .., N$

Where ω is the weight vector, b is the bias, and $\Phi(x_i)$ is the high dimensional feature space. $C < 0$ is a prespecified constant that is responsible for regularization and represents the weight of the loss function. The first term of the objective function $\omega^T \omega$ is the regularized term and the second term $C \sum_{i=1}^{N} (\zeta_i + \zeta_i^*)$ is called the empirical term and measures the $\varepsilon-$insensitive loss function. ζ_i and ζ_i^* are the slack variables to guard against outliers, they represent the distance between the potential support vector and the potential outliers.

Kernel methods achieve flexibility by fitting simple models in a local region to the target point x. Localization is achieved via a weighting kernel K, and individual observations receive weights $K(x_i, x)$, so the Radial basis function is written as:

$$f(x) = \sum_{i \in SV} (\alpha_i - \alpha_i^*) K(x_i, x) \tag{22}$$

With $K(x_i, x) = \exp(-\gamma \,||\, x_i - x \,||^2)$

$\alpha_i - \alpha_i^*$ are coefficients of the support vector in the decision function, γ is the kernel coefficient and x is the center of feature.

Decision Trees Regression (DTR)

DTR uses a tree-like structure to model the relationship between the set of predictors and the response. The tree is composed of nodes that represent the possible values of the predictors or the response. Its purpose is to find the best split at each node. The quality of a candidate split θ of node m is then computed using an impurity function or loss function $H(Q_m)$, the choice depends on the task being solved:

$$G(Q_m, \theta) = \frac{n_m^{left}}{n_m} H(Q_m^{left}(\theta)) + \frac{n_m^{right}}{n_m} H(Q_m^{right}(\theta)) \tag{23}$$

Select the parameters that minimizes the impurity:

$$\theta^* = argmin_\theta G(Q_m, \theta) \tag{24}$$

Q_m is the data at node m partitioned into Q_m^{left} and Q_m^{right} with n_m sample. $H(Q_m)$ is the Mean Squared Error (MSE) such as ($\bar{y}_m$ is the mean of predicted value):

$$H(Q_m) = \frac{1}{n_m} \sum_{y \in Q_m} (y - \bar{y}_m)^2 \quad (25)$$

Ridge Regression (RR)

RR shrinks the Least Squares unbiased coefficients by imposing a penalty on their size. It produces biased estimators of regression coefficients, that have a small variance and more stable than the Least Squares unbiased coefficients, which called ridge estimators. The coefficients minimize a penalized residuals sum of square. For a given value of λ, a non-negative parameter, RR solves the problem [ROD22]:

$$min_{(\beta_0 \beta)} \left(\frac{1}{2N} \sum_{i=1}^{N} (y_i - \beta_0 - x_i^\top \beta)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right) \quad (26)$$

where N and p are the observation and predictor numbers, respectively.

Partial Least Squares Regression (PLSR)

PLS regression constructs a set of latent variables. These latent variables are linear combinations of the original predictors, created in such a way that they explain the maximum covariance between the predictors and the response variable. PLSR model with h latent variables can be expressed as follows [CHO05]:

$$X = TP^t + E \quad (27)$$

$$y = Tb + f \quad (28)$$

In Equation (27, 28) $X(n \times p), T(n \times h), P(p \times h), y(n \times 1)$ and $b(h \times 1)$ are respectively used for predictors, X scores, X loadings, response, and regression coefficients of T. The $k-th$ element of column vector b explains the relation between y and t_k, the $k-th$ column vector of T. Meanwhile, $E(n \times p)$ and $f(n \times 1)$ stand for random errors of X and y, respectively.

4 EXPERIMENTS AND DATA SETS

4.1 Experiments

4.1.1 Sample preparation

The experiments were conducted on well-fermented commercial cocoa beans from the same harvest south of Ivory Coast. We extracted $15kg$ from $25kg$ fermented cocoa beans on the last day of fermentation (day 7). Once in the laboratory, the beans are distributed in batches of 10, 30, 50 and 70 in polystyrene bags, then stored in a cold room at $-10°C$ throughout the handling process. Before using, the cocoa beans are defrosted at room temperature in the laboratory.

4.1.2 Experimental process

The experimental process is composed of three steps: drying, image acquisition, and weight determination, as shown in the experimental cycle (figure 3).

- *Cocoa drying*: the cocoa bean batches were dried in a domestic microwave oven (SHARP R-75 MT). Intermittence drying is used with a pulsing ratio (Equation 29) of 4 to limit local overheating and a drying power of $270W$. This means 2 minutes of microwave start-up and 4 minutes of shutdown. Drying was stopped when the moisture content reached around 7%. Data are acquired at regular time intervals during drying, this resulted in 30 to 45 observations per batch

- *Weight determination*: the weight of the beans batch is determined by a digital precision scale with 0.01g precision.

- *Image acquisition*: A setup for capturing customized images has been developed. It is composed of: *a color coupled charge device (CCD) camera* (SONY XCG-5005CR, Japan), which is specifically standardized for machine-vision applications based on Gigabit Ethernet technology, *a lens zoom 16 mm* (Fujifilm corporation, model HF16HA-1B, Japan). *The image acquisition card* (Mil Matrox) is used for transferring information from camera to computer (Core-i7 CPU: 2.5 GHz; RAM: 4 GB). And two 8.5 watts white LED, which cover the visible wavelength, to ensure correct and consistent lighting throughout the acquisition process. The set is placed in a closed box to control the lighting. The images have been acquired in tiff format, 2448 x 2048 definition, 96 ppi (horizontal) and 96 ppi (vertical) and unit 8.

figure 4 shows some cocoa beans images at different times with their moisture content.

$$PR = \frac{CyclePoweronTime + CyclePoweroffTime}{CyclePoweronTime} \quad (29)$$

4.1.3 Moisture content computation

At the end of the drying process, the dry mass of each batch is determined by drying the dried beans in an oven at $103°C$ for 16 hours to determine the dry weight. The weights determined during drying are used to calculate the water content using the following formula (ISO 2451/2014 standard):

$$MC_t\ (wet\ base) = \frac{m_t - m_s}{m_t} \quad (30)$$

With MC the moisture content, m_t the weight at time t and m_s the dry weight.

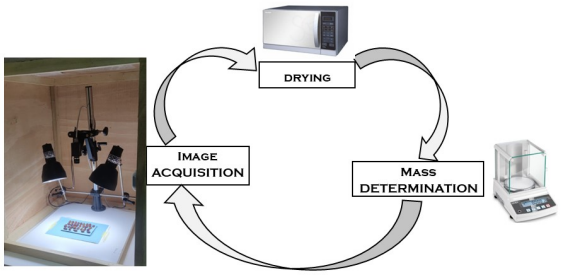


Figure 3: experimental cycle.

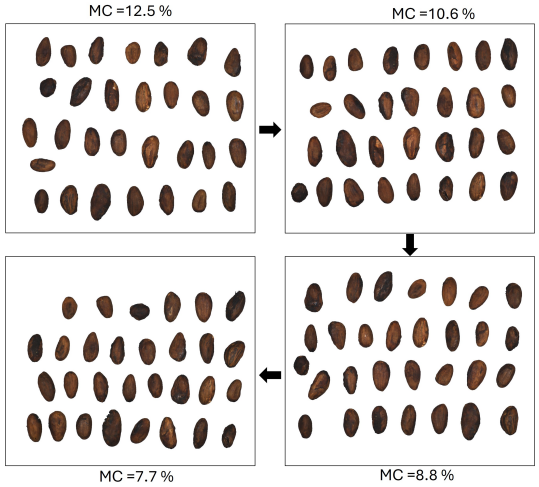


Figure 4: Cocoa bean images as a function of MC

4.2 Dataset description

The data come from 4 different samples, depending on the number of beans per batch or per image. E10, E30, E50 and E70 samples, with 10, 30, 50 and 70 beans per batch respectively. Each sample consists of 5 batches. For the variable selection, the dataset of a sample is made up of all the observations of the 5 batches of this sample. For the model analysis, the dataset consists of 80% of the dataset for training and 20% for test. The dataset consists of color features from segmented images, as predictors and moisture content as response. Application of extraction methods to the color components yielded 90 candidate variables. The database for each sample is then standardized using the z-score method. The Z-score standardization involves transforming each feature in the dataset such that it has a mean 0 with a unit standard deviation (Equation 31).

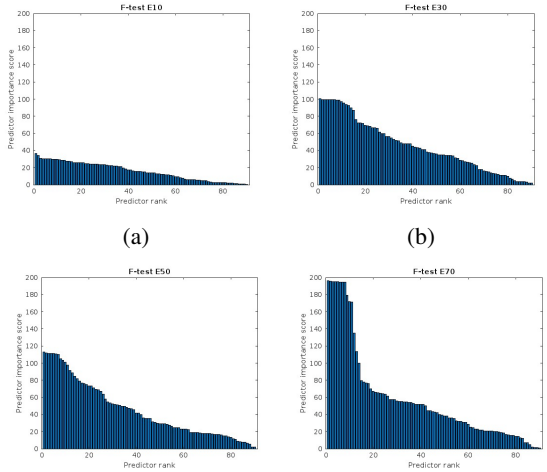
$$z = \frac{(x-\mu)}{\sigma} \tag{31}$$

Where μ, σ and x are the mean, the standard deviation and the feature value of the original dataset.

5 RESULTS AND DISCUSSION

5.1 Features extraction and selection

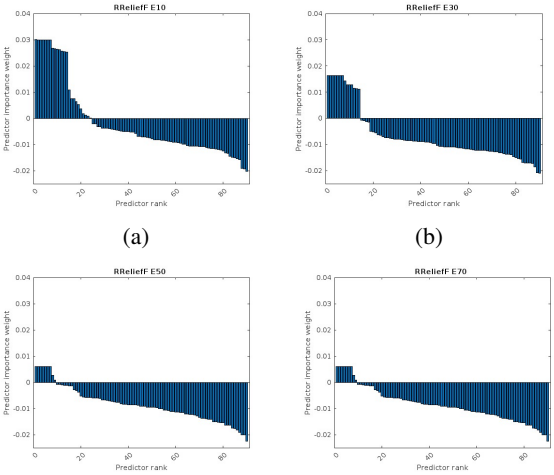
Figure 5a to figure 6d show the importance scores of each candidate variable for each sample, which (a), (b), (c) and (d) correspond to samples E10, E30, E50 and E70, respectively. The variables are ranked in importance order.



(a) (b)

(c) (d)

Figure 5: F-test predictor importance score



(a) (b)

(c) (d)

Figure 6: RReliefF predictor importance score

Figure 5a to Figure 5d display the significant score for the F-test method. The predictors importance score maximums are 36.36, 100.36, 113.10 and 196.30 from respectively samples E10, E30, E50 and E70. The significance of each variable increases as the number of beans per image increases, and the sample E70 gives the best scores. Sample E70 allows to distinguish the most significant predictors. The most significant features are those that stand out from the others in the E70 sample. Their scores range from 99.73 to 196.30. There are

14 of them, with scores ranging from 99.73 to 196.30. Even if samples $E30$ and $E50$ don't help to distinguish them, they are still ranked in the same order, except in sample $E10$.

Figure 6a to Figure 6d show the RReliefF predictor importance weight. The weights reflect how each feature discriminates between instances of different classes or categories in the dataset. Unlike F-test, the predictor importance score don't increase as the number of beans per image increases. Although the weights are different, the 4 samples display the same list of the most relevant features. The most relevant features are those that have positive weights.

Using both methods and based on several experiments, a feature is said to be relevant when its F-test score is greater than 55 and its RReliefF weight is positive. Thus, the relevant features for predicting moisture content are energy and entropy of rg, yb, Cb, Cr, a^*, b^* and h^* components. This result is confirmed by the four samples. Of these features, energy-based variables are better than entropy-based features. On can also observe that the energy-based features have approximately the same scores in the F-test and RReliefF.

As a reminder, the database used for feature selection is made up of a set of observations from several different batches of beans. Thus, the redundancy of features can be explained by poor correlation with the moisture content or by the instability of features. In the case of instability, the feature may correlate well with the moisture content for a given batch. However, the values of this feature vary from batch to batch. The poor results for the features selected for samples E10 and E30 can be explained by the wide range of cocoa bean colors. These samples don't contain enough beans to take into account the maximum colors of cocoa bean.

The almost identical scores of energy-based features on the one hand and entropy-based features on the other may be due to the multi-collinearity between features. To confirm this hypothesis, we calculated the Variance Inflation Factors (VIF) of each selected feature by equation (32). VIF values range from 57.54 to $1.1x10^{+6}$. These values, being well above 10, show that the selected features are highly correlated [MONT21].

$$VIF_j = \frac{1}{1-R_j^2} \tag{32}$$

where R_j^2 is the coefficient of multiple determination obtained from regressing predictor x_j on other predictors. As the selected variables are highly correlated, if they are directly included in the models, this can create instability and over-fitting model due to the inflation of regression coefficients. Thus, for the model analysis in the next paragraph, PCA is used to select decorrelated variables.

5.2 Machine Learning models

The data is randomly divided into training and testing sets 20 times during the 20 training sessions of the models. The models are analyzed on two sets of predictors. The first set is made up of all selected predictors, and the second set is made up of energy-based predictors. The scikit-learn library in Python is used for this analysis. Coefficient of determination (R^2) and Root Mean Square Error $(RMSE)$ are the metrics used to evaluate models, by calculating the mean and SD of the five datasets. The hyper-parameters of each model are found with the Grid Search Cross-Validation (GridSearchCV). The hyper-parameters of SVR/RBF are regularization parameter C =10, $\gamma = 0.055$ and $\varepsilon = 0.01$. Ridge trace using all predictors gives ridge parameter $\alpha = 0.001$. For all selected predictors, there are 8 principal components for MLR, SVR, DTR, and PLSR models, and 35 principal components for the polynomial model with interaction. For the energy-only predictors, 5 principal components for MLR, SVR, DTR, and PLSR models, and 30 principal components for the polynomial model with interaction.

Table 1 and table 2 show the standard deviation (SD) and the mean of evaluation metrics for the models. SD is used to assess the stability of models. Concerning model training with all selected predictors (Table 1), more than 94% of proportion of variance in the MC is predicted from the predictors, for polynomial, SVR and DTR models. These models have mean errors of less than 5% of MC. On the other hand, for Ridge and PLSR models, less than 90% of variability in the training MC explained by the models and have mean errors of more than 5% of MC. The standard deviations of coefficient of determination for all models are less than 1%, indicating their stability during training. When testing the models with all selected predictors (Table 1), less than 90% of variability in the training MC explained by the models and less than 6% of mean errors of MC. There is a significant difference between the training and test of Polynomial, SVR, and DTR models, but this is not the case for Ridge and PLSR models. We also note that the SDs of testing are larger, more than 2% for the coefficient determination. The same observation for models trained with the energy features. However, the predictions with all selected predictors are better than energy predictors. All the models studied, for both sets of predictors, have results far better than Hashimoto et al, who obtained a coefficient of determination of 0.67 by predicting water content with near-infrared diffuse reflectance spectra [HII09]. These results are still inferior to those obtained using drying time as a predictor [IGO15], [KAR18].

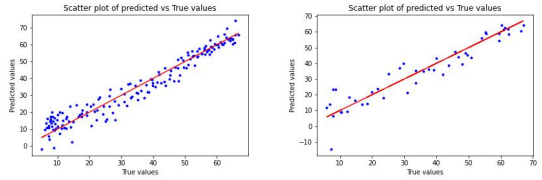
Figure 7a to Figure 11b displays the moisture content predicted value by the model versus the true value. The training and test data fit well on the straight line for the

		$R^2_{training}$		$R^2_{testing}$		$RMSE_{training}$		$RMSE_{testing}$	
Models	hyperparameters	*Mean*	*SD*	*Mean*	*SD*	*Mean*	*SD*	*Mean*	*SD*
PCA + Polynomial	$degree = 2$	0.94	0.004	0.83	0.11	4.81	0.2	7.59	2.20
PCA + SVR	rbf	0.95	0.003	0.90	0.02	4.54	0.16	6.02	0.66
PCA + DTR	$depth = 7, min-samp.-spl. = 15$	0.96	0.004	0.86	0.07	3.85	0.22	7.00	1.83
Ridge R.	$alpha = 0.001$	0.89	0.01	0.86	0.08	6.50	0.43	7.01	1.66
PLSR	$component = 8$	0.86	0.01	0.83	0.09	7.24	0.41	7.70	1.90

Table 1: results of training and testing models for all selected predictors

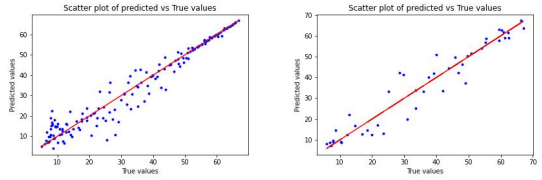
		$R^2_{training}$		$R^2_{testing}$		$RMSE_{training}$		$RMSE_{testing}$	
Models	hyperparameters	*Mean*	*SD*	*Mean*	*SD*	*Mean*	*SD*	*Mean*	*SD*
PCA + Polynomial	$degree = 2$	0.93	0.005	0.82	0.10	5.28	0.20	7.73	2.15
PCA + SVR	rbf	0.88	0.01	0.87	0.05	6.72	0.33	6.80	1.38
PCA + DTR	$depth = 5$, $min-samp.-spl. = 10$	0.94	0.004	0.83	0.07	4.59	0.20	7.59	1.62
Ridge R.	$alpha = 0.001$	0.86	0.01	0.84	0.07	7.20	0.55	7.36	1.53
PLSR	$component = 5$	0.85	0.01	0.84	0.07	7.52	0.45	7.54	1.60

Table 2: results of training and testing models for energy predictors



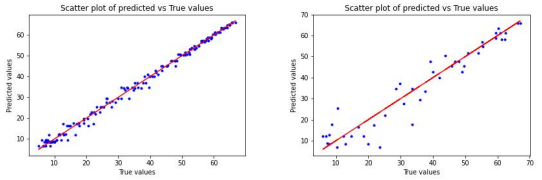
(a) (b)

Figure 7: Scatter plot of predicted vs True values for PCA + Polynomial model for all selected predictors; (a) training (b) testing



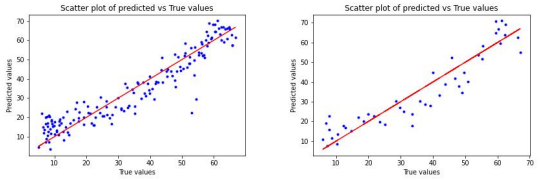
(a) (b)

Figure 8: Scatter plot of predicted vs True values for PCA + SVR model for all selected predictors; (a) training (b) testing



(a) (b)

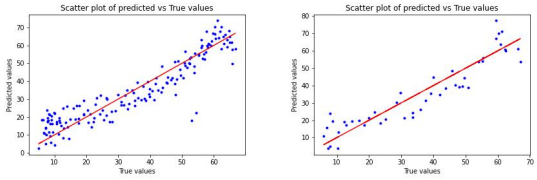
Figure 9: Scatter plot of predicted vs True values for PCA + DTR model for all selected predictors; (a) training (b) testing



(a) (b)

Figure 10: Scatter plot of predicted vs True values for RIDGE model for all selected predictors; (a) training (b) testing

non-linear models, Polynomial, SVR, and DTR (Figure 7a to Figure 9b. The linear models, Ridge and PLSR reveal outliers and significant deviations in the 30 to 50 MC range (Figure 10a to Figure 11b). This shows that the evolution of MC during drying is not linear. The tree-like structure is more effective in fitting all the training data as compared to the other two non-linear structures. It is important to note that the model performs well for MC values above 45%, while the fit is slightly less for values below 45%. Both the training and the test data show good results for the given model. The ideal moisture content for declaring cocoa to be dry is 7-8%. Thus, the most important MC range is below 10%. Support Vector Regression predicts 95% of training moisture content, 90% of new moisture content, and well the moisture content below 10%. The standard deviation shows that this model is more stable than others. Therefore, For this particular study, it has shown the most promising results in predicting the moisture content of cocoa beans during the drying process. The model uses predictors based on the energy and entropy of rg, yb, Cb, Cr, a*, b*, and h* components. However, it's important to note that the mean error still high in predicting new data.



(a) (b)

Figure 11: Scatter plot of predicted vs True values for PLSR model for all selected predictors; (a) training (b) testing

6 CONCLUSION

This article aims to analyze the color features of unshelled cocoa beans during the drying process. Additionally, it studies various linear and non-linear Machine Learning models to predict moisture content based on color features. The article analyzes the mean, standard deviation, entropy, energy, kurtosis, and skewness of the components of the RGB, YCbCr, HSV, and Lab color spaces, without luminance components, using both the F-test and RReliefF methods. The analysis is performed on samples of 10, 30, 50, and 70 beans per batch. The color components that are most relevant for predicting moisture content during drying are derived from the energy and entropy of YCbCr, and Lab color spaces. The relevance of these components becomes more important as the number of beans in a batch increases. Additionally, the selected features are highly interdependent. Non-linear models provide more accurate moisture content predictions during drying than linear models, precisely Support Vector Regression with radial basis function performs better. To conclude, moisture content can be predicted during drying with color image features. This article opens the way for the study of cocoa beans moisture content prediction using image data. It provides relevant information on the evolution of water content during drying as a function of colour characteristics and also in the different colour spaces.

7 ACKNOWLEDGMENTS

This work is financed in part by Ministère des Affaires Etrangères - France from Service de Coopération et d'Action Culturelle (SCAC) of Embassy of France in Ivory Coast.

8 REFERENCES

[CAS23] E. Castillo-Orozco, O. Garavitto, O. Saavedra, et D. Mantilla, The Drying Kinetics and CFD Multidomain Model of Cocoa Bean Variety CCN51 , Foods, vol. 12, no 5, p. 1082, mars 2023, doi: 10.3390/foods12051082.

[CAS20] M. Castelli, F. M. Clemente, A. Popovic, S. Silva, et L. Vanneschi, A Machine Learning Approach to Predict Air Quality in California , Complexity, vol. 2020, p. e8049504, August 2020, doi: 10.1155/2020/8049504.

[CHO05] I.-G. Chong et C.-H. Jun, Performance of some variable selection methods when multicollinearity is present , Chemom. Intell. Lab. Syst., vol. 78, no 1-2, p. 103-112, juill. 2005, doi: 10.1016/j.chemolab.2004.12.011.

[DJE09] A. Djedjro, E. Assidjo, K. Patrice, et B. Yao. Mathematical Modelling of Sun Drying Kinetics of Thin Layer Cocoa (Theobroma Cacao) Beans, J. Appl. Sci. Res., vol. 5, p. 1110-1116, sept. 2009.

[HAS18] J. C. Hashimoto et al. Quality Control of Commercial Cocoa Beans (Theobroma cacao L.) by Near-infrared Spectroscopy , Food Anal. Methods, vol. 11, no 5, p. 1510-1517, mai 2018, doi: 10.1007/s12161-017-1137-2.

[HII09] C. L. Hii, C. L. Law, M. Cloke, et S. Suzannah. Thin layer drying kinetics of cocoa and dried product quality , Biosyst. Eng., vol. 102, no 2, Art. no 2, fÃ©vr. 2009, doi: 10.1016/j.biosystemseng.2008.10.007.

[HII11] C. L. Hii, C. L. Law, et S. Suzannah. Drying kinetics of the individual layer of cocoa beans during heat pump drying, J. Food Eng., vol. 108, no 2, p. 276-282, janv. 2012, doi: 10.1016/j.jfoodeng.2011.08.017.

[HUM10] E. M. Humston, J. D. Knowles, A. McShea, et R. E. Synovec. Quantitative assessment of moisture damage for cacao bean quality using two-dimensional gas chromatography combined with time-of-flight mass spectrometry and chemometrics, J. Chromatogr. A, vol. 1217, no 12, Art. no 12, mars 2010, doi: 10.1016/j.chroma.2010.01.069.

[IGO15] N. A. L. Igor, A. D. Clement, B. Kouakou, et A. N. Emmanuel, Modélisation de la cinétique de séchage des feves de cacao par des modeles semi-empiriques et par un réseau de neurones artificiels récurrent: cas du séchage microonde par intermittence, p. 16, 2015.

[KAR18] D. Karidioula, D. C. Akmel, N. E. Assidjo, et A. Trokourey, Modélisation du séchage solaire de feves de cacao par le Réseau de Neurones Artificiel , Int. J. Biol. Chem. Sci., vol. 12, no 1, p. 195, juin 2018, doi: 10.4314/ijbcs.v12i1.15.

[KAV21] M. Kaveh, R. Chayjan, I. Golpour, S. Poncet, F. Seirafi, et B. Khezri, Evaluation of exergy performance and onion drying properties in a multistage semi-industrial continuous dryer: Artificial Neural Networks (ANNs) and ANFIS models ,

Food Bioprod. Process., vol. 127, p. 58-76, February. 2021, doi: 10.1016/j.fbp.2021.02.010.

[MATFT24] Univariate feature ranking for regression using F-tests - MATLAB fsrftest - MathWorks France . Accessed: February 25, 2024. [Online]. Available: https://fr.mathworks.com/help/stats/fsrftest.html

[MATSG24] Segment Image and Create Mask Using Color Thresholder - MATLAB and Simulink - MathWorks France. Accessed: May 07, 2024. [Online]. Available: https://fr.mathworks.com/help/images/image-segmentation-using-the-color-thesholder-app.html

[MONT21] D. C. Montgomery, E. A. Peck, et G. G. Vining, Introduction to Linear Regression Analysis. John Wiley and Sons, 2021.

[OME14] N. Omer Fadl Elssied, O. Ibrahim, et A. Hamza Osman, A Novel Feature Selection Based on One-Way ANOVA F-Test for E-Mail Spam Classification , Res.J. Appl. Sci. Eng. Technol., vol. 7, no 3, p. 625-638, janv. 2014, doi: 10.19026/rjaset.7.299.

[RAS06] Color Image Processing Methods and Applications by Rastislav Lukac, Kostantinos N. Plataniotis

[ROB97] M. Robnik-Sikonja et I. Kononenko, An adaptation of Relief for attribute estimation in regression , presented in International Conference on Machine Learning, juill. 1997.

[ROD22] A. Rodríguez Sánchez, R. Salmerón Gómez, et C. García, The coefficient of determination in the ridge regression , Commun. Stat. - Simul. Comput., vol. 51, no 1, p. 201-219, janv. 2022, doi: 10.1080/03610918.2019.1649421.

[WAN14] X.-Y. Wang, B.-B. Zhang, et H.-Y. Yang, Content-based image retrieval by integrating color and texture features , Multimed. Tools Appl., vol. 68, no 3, p. 545-569, February. 2014, doi: 10.1007/s11042-012-1055-7.