Computer Vision Based Detection and Quantification of Extraneous Water in Raw Milk

Bezuayehu Gutema Asefa (bezuayehug7@gmail.com)
Ethiopian Institute of Agricultural Research https://orcid.org/0000-0002-1525-2054
Legesse Hagos
Ethiopian Institute of Agricultural Research
Tamirat Kore
Ethiopian Institute of Agricultural Research
Shimelis Admassu Emire
Addis Ababa University

Research Article

Keywords: Milk adulteration, Multivariate classification, Support vector machine, Validation

Posted Date: June 29th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-625039/v1

Abstract

A rapid method based on digital image analysis and machine learning technique is proposed for the detection of milk adulteration with water. Several machine learning algorithms were compared, and SVM performed best with 89.48 % of total accuracy and 95.10 % precision. An increase in the classification performance was observed in extreme classes. Better quantitative determination of the extraneous water was achieved using SVMR with R^2(CV) and R^2(P) of 0.65 and 0.71 respectively. The proposed technique can be used to screen raw milk based on the level of added extraneous water without the necessity of any additional reagent.

1. Introduction

With high nutritive value, providing macronutrients (proteins, fat and minerals) and micronutrients (vitamins and trace elements), cow milk is among the recognized contributor to a balanced diet of many populations. Due to the high nutritional composition, a high rate of milk consumption with an increasing demand exists worldwide (Handford et al., 2016). Despite the role of milk in food and nutrition security, the increase in demand has amplified fraudulent activities, subsequently making milk the second most vulnerable product to adulteration (Moore et al., 2012).

Milk adulteration could be dilution with water with the intention to increase economic gain or addition of substances (*e.g.*, Sucrose, sodium chloride, vegetable oil and surfactants) that improve the physicochemical and visual characteristics of milk (Poonia et al., 2017). Besides, the addition of substances that extend the shelf life of milk, such as formaldehyde, hydrogen peroxide and hypochlorite is becoming a serious issue of adulteration in the dairy industry (Das et al., 2016; Handford et al., 2016; Nascimento et al., 2017).

Assessments on the prevalence of milk adulteration in several countries found water as the most frequently added adulterants (Faraz et al., 2013; Kandpal et al., 2012; Shabir Barham, 2014; Soomro et al., 2014). Water is added to grow economic gain by increasing the volume of milk through dilution. However, the addition of water to milk dilutes the constituents in milk and could cause potential public health risk of acute malnutrition (stunting, wasting and underweight) which leads to nutrition-related child mortality (Handford et al., 2016; Park et al., 2013; Shabir Barham, 2014). According to experts, next to educating farmers about the consequences of milk fraud, the need for improved detection is key to address the prevailing risk of fraud in milk (Handford et al., 2016).

Several studies have shown the possibility of determining the presence of water as an adulterant in milk samples using different techniques. Newly developing techniques that are robust, green, simple and cost-effective are gaining increasing importance in food quality monitoring. Digital image-based procedures that use the power of machine learning algorithms are increasingly used to assess adulteration in agro-food products including milk. In recent years, several studies were conducted to develop digital image-based techniques for the determination of adulterants in milk. However, the newly developed techniques lack representative sampling during imaging of milk samples. In addition, indicator chemicals were used to bring the desired classification result before the imaging process (Dos Santos & Pereira-Filho, 2013; Kobek, 2017). This brings limitations in the utilization of those techniques since users of such methods are required to have technical knowledge of the procedure.

Considering the limitation in the existing methods, this paper proposed a clean method based on digital image processing coupled with a machine-learning algorithm to test milk adulteration with water. The proposed technique is fast, robust and doesn't require sample preparation and use of any chemicals.

2. Materials And Methods

2.1. Milk Samples

Raw milk samples were obtained from two different dairy farms found in Sebeta and Debre Zeit Agricultural Research Centers of the Ethiopian Institute of Agricultural Research (EIAR). Known research dairy farms were selected to ensure the purity of the milk before spiking the adulterant. A batch of milk was used to acquire images of pure milk and modified milk with water as an adulterant in a range from 10 to 40%. Since image acquisition was performed in sampling locations, all milk samples used in the study were neither refrigerated nor subjected to transportation longer than one km. The volume of milk sample for each image acquisition was kept constant at 25 ml, which was quantitatively transferred to a petri dish to acquire images from the top surface. Adulterated milk samples were simulated by spiking water in the whole sample used in one day to avoid differences in image intensities due to spiking individual samples.

2.2. Image Acquisition

A conventional image acquisition chamber having a dimension *L x W x H* of *(40 x 40 x 60)*, made from aluminum sheet was used. Uniform lighting was maintained using twelve fluorescent lumps mounted to four sides of the imaging chamber at a height 40 cm above the bottom surface. A digital camera (EOS, 6D Mark II, Canon, Japan) installed with an image stabilizer of 24–105 mm was set at the top of the image acquisition chamber heading down to the petri dish containing milk sample at a height of around 55 cm. The process of image acquisition was fully monitored using EOS utility software. Fifty samples were prepared for each sample group from the two sampling sites. Image of each sample was captured in duplicate, making a total of two hundred images for each group of samples.

2.3. Feature extraction

Images acquired from all samples were processed using a batch processor in ImageJ. All the captured images were treated with a global processing stage, that takes the region of interest from the bulk image. The central area of each image was cropped with a pixel size of 250 x 250. Further processing such as converting to different color spaces (Lab* and HSI) and filtering were performed as summarized in (Table 1). The mean and modal grey values, minimum and maximum grey values, standard deviation, median and center of mass were calculated for each processed image. After calculating processed image parameters, some values indicated in the '-' sign in (Table 1) are found irrelevant and were not included as a predictor variable due to the fact that similar output values were obtained for all sample groups. Totally, 125 variables were included as a predictor in the development of multivariate models.

Table 1
Description of the image processing and parameters included as a variable for the development of classification model

Image process description	Measurement parameters									
	Mean grey value	Standard deviation	Modal grey value	Median grey value	Minimum grey value	Maximum grey value	Center of mass (X maximum)	Center of mass (Y maximum)	Skewness	Kurtosis
Resizing (250 x 250 pixels)	+	+	+	+	+	+	+	+	+	+
Filtering (Gaussian)	+	+	+	+	+	+	+	+	+	+
Filtering (Median)	+	+	+	+	+	+	+	+	+	+
Filtering (Kwahara)	+	+	+	+	+	+	+	+	+	+
Filtering (FFT)	+	+	+	+	-	-	+	+	+	+
Filtering (Convolve)	+	+	-	-	-	+	+	+	+	+
Splitting RGB (R)	+	+	+	+	+	+	+	+	+	+
Splitting RGB (G)	+	+	+	+	+	+	+	+	+	+
Splitting RGB (B)	+	+	+	+	+	+	+	+	+	+
Convert to HSI (H)	-	-	-	-	-	-	-	-	-	-
Convert to HSI (S)	-	-	-	-	-	-	-	-	-	-
Convert to HSI (I)	+	+	+	+	+	+	+	+	+	+
Convert to Lab* (L)	+	+	+	+	+	+	+	+	+	+
Convert to Lab* (a*)	+	+	+	+	+	+	+	+	+	+
Convert to Lab* (b*)	+	+	+	+	+	+	+	+	+	+

Image processing description: '+' signs indicate parameters used as a variable, whereas '-' signs refer to parameters excluded from the variable list

2.4. Multivariate procedure

Numerical values generated from the processed images were used to develop classification and regression models based on the level of added water into the pure milk. Multivariate procedures were Performed using MATLAB software (R2020b, PLS Toolbox, Eigenvector). Characteristics of the different multivariate procedures used in the current study are briefly described in Table (2).

Table 2: Summary of machine learning algorithms used for the classification task

Table 2
Summary of machine learning algorithms used for the classification task

Algorithm	Description
K-nearest neighbor (K-NN)	K-NN-based classification works by identifying the distances between an unknown object and each of the objects of the training set mostly based on the Euclidean distance. A decision is made based on the majority rule after the selection of the k-nearest objects to the unknown sample (Berrueta et al., 2007).
Soft independent modeling of class analogy (SIMCA)	SIMCA calculates the geometric distance from the principal component model and determines the class distance. In addition, the modeling and discriminatory powers are determined (Brereton, 2003).
Support vector machine (SVM)	SVM-based classification works by obtaining the 'optimal' boundary of two classes in a vector space independently on the probabilistic distributions of training vectors in the data set (Berrueta et al., 2007).
Partial least square discriminant analysis (PLS-DA)	PLS-based classification works by finding the components in the input matrix (X) that describe the relevant variations at most in the input variables and have a maximal correlation with the target value in Y (Massart et al., 1998).

2.5. Model performance evaluation

The performance of each model was assessed using a total accuracy method which was computed using the True Positive (TP) and True Negative (TN) values obtained from the confusion matrix (Eq. 1) (Tang et al., 2014). Besides, the precision (Eq. 2) recall (Eq. 3) was calculated based on False Negative (FN) and False Positive (FP) values to support the classification model effectiveness (Lopes et al., 2019).

$$Accuracy = TP/(TP + FN) \quad (1)$$

$$Precision = TP/(TP + FP) \quad (2)$$

$$Recall = TP/(TP + FN) \quad (3)$$

3. Results

3.1. Exploratory Analysis

A total of 25 predictor variables from 900 image data *(i.e., 180 x 5 groups)* were inspected visually from the excel file to identify potential outliers. Based on the observation, 29 image data were removed and the remaining 871 image data were used to develop the classification models. Before the analysis, Kenard stone technique was employed to randomly separate 80% of the data into the training set and the remaining 20% into a test set. The effect of variation in feature size was corrected by autoscaling the predictor variables.

Principal Component Analysis (PCA) was applied to reduce data dimensionality and new variables that are linear combinations of the original image feature values were generated. The selection of an optimal number of PCs was done based on the lowest prediction error in cross-validation (Venetian blinds).

3.2. Multivariate Classification

The result table indicating the performance of each classification algorithm is given in Table (3). Of the four classification algorithms, SIMCA provided the worst performance with less than 60% total accuracy in a training dataset. Next to SIMCA, poor classification performance was obtained with the PLSDA algorithm. In contrast to the two classifiers, KNN and SVM achieved fair classification with total accuracy of 79.45 and 89.48 respectively. SVM generally achieved superior results compared to all the classifiers with 89.48% accuracy, 95.10% precision, and 83.24% recall values.

Table 3: Performance measures of different classification algorithms over the training, cross-validation and prediction dataset.

Table 3
Performance measures of different classification algorithms over the training, cross-validation and prediction dataset.

Algorithm	Performance measures	Training set	Cross-validation set	Testing set
KNN	Accuracy	79.43	81.78	79.45
	Precision	89.22	90.35	88.46
	Recall	67.16	71.12	67.59
SVM	Accuracy	100	86.47	89.48
	Precision	100	93.38	95.10
	Recall	100	78.58	83.24
PLS-DA	Accuracy	66.94	66.12	66.97
	Precision	75.91	74.39	75.98
	Recall	48.07	46.72	47.90
SIMCA	Accuracy	58.49		
	Precision	41.50		
	Recall	87.12		

Table 4: Performance measures for class prediction of KNN, PLS-DA and SVM algorithms

Table 4
Performance measures for class prediction of KNN, PLS-DA and SVM algorithms

Algorithm	M: W	Training	Cross-validation	Testing
KNN	0%	81.91	81.95	78.04
	10%	78.13	79.93	79.38
	20%	76.75	80.85	78.59
	30%	76.43	77.88	72.04
	40%	83.95	88.28	89.19
PLS-DA	0%	63.98	63.75	68.74
	10%	67.13	66.56	67.57
	20%	58.31	56.58	62.73
	30%	56.94	55.82	55.48
	40%	88.37	87.88	80.35
SVM	0%	100	87.07	91.95
	10%	100	85.27	88.26
	20%	100	84.72	88.99
	30%	100	82.82	86.18
	40%	100	92.47	92.04

3.3. Estimation of adulteration level

The dataset was also used to develop a prediction model for the level of adulteration. The prediction performance of Partial Least Squares Regression (PLSR), Principal Component Regression (PCR), and SVMR algorithms was evaluated. The summary of quantitative prediction performance measures is presented in Table (5).

Table 5: Performance measures of regression models developed for quantitative adulterant prediction

Table 5
Performance measures of regression models developed for quantitative adulterant prediction

No.	Method	preprocess.	LV/PC	RMSEC	RMSECV	R^2(Cal)	R^2(CV)	R^2(P)
2	PCR	>>	3	11.23	11.28	0.31	0.30	0.16
3	PLSR	>>	6	9.84	10.05	0.47	0.44	0.44
4	SVMR	>>		4.93	8.02	0.87	0.65	0.71

4. Discussion

The exploratory analysis showed that the first three PCs explained more than 75% of the data variance as indicated in a 3-dimensional PCA score-plot obtained from three PCs (Fig. 1). The change in color intensity can be observed from the score-plot. Increasing the amount of added water could be related to the diminishing color density of the images which is illustrated in reduced scores in PC 1. Since milk color is influenced by the composition, the addition of water to pure milk can affect the intensity. Detecting such minor differences in the intensity of milk color using the human eye could be difficult unless digital technologies are used with the support of numerical software.

Further analysis on the model's prediction performance for each class of samples exhibited efficient classification performance of SVM algorithms in extreme classes Table (4). This means milk samples with no adulteration and milk samples that have 40 % added water were identified with better classification performance compared to other samples. Correct identification of pure milk sample was achieved using the same algorithm with an accuracy of 91.95% in prediction set samples. Also, SVM achieved the highest classification accuracy (92.04) in milk samples adulterated with 40% water.

This result outperformed the previously developed procedure (Kobek, 2017), who found total classification accuracy of 81.66 using an Artificial Neural Network (ANN) based classification model. In another research, SIMCA and KNN classification algorithms were applied to distinguish milk adulterated with water from pure milk, and total accuracy of 82 and 92% respectively for SIMCA and KNN were found (Dos Santos & Pereira-Filho, 2013). However, indicator chemicals were used to bring the desired color change in the two findings. Given these facts, our finding verified the possibility of using digital images to determine milk adulteration with water without the necessity of adding indicator chemicals.

Except for the SVMR algorithm, inadequate prediction performance was found in predicting the level of extraneous water with prediction R^2 of 0.16, 0.44 and 0.52 in PCR, PLSR and MLR respectively. Interestingly, SVMR achieved better performance in predicting the amount of adulterated water in the milk samples with R^2(CV) and R^2(P) of 0.65 and 0.71 respectively.

5. Conclusion

The change in color of milk due to dilution by water has proved to be useful to detect adulteration through the use of processed images coupled with machine learning algorithms. SVM classification model discriminated milk samples based on the level of added water with accuracy and precision of 89.48 % and 95.10%, respectively. The performance of the proposed technique is satisfactory to apply for screening of raw milk samples at dairy processing industry. The proposed technique can be used for the rapid determination of extraneous water in raw milk without the necessity of any additional reagent.

Abbreviations

SVM: Support Vector Machine;

SVMR: Support Vector Machine Regression;

CV: Cross validation;

P: Prediction;

HIS: Hue, Intensity, Saturation;

TP: True positive;

TN: True negative;

FN: False negative

FP: False positive

PCA: Principal Component analysis

SIMCA: Soft Independent Modeling of Class Analogies

KNN: K- nearest neighbors

PLSDA: Partial Least Squares Discriminant Analysis

PCR: Principal Component Regression

PLS: Partial Least Squares

PLSR: Partial Least Square Regression

EIAR: Ethiopian Institute of agricultural Research

Declarations

Ethical approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

Data can be shared upon request

Competing interests

The authors declare that they have no known competing interests.

Funding

Not applicable

Authors' contributions

Bezuayehu G. Asefa: Conceptualization, Methodology, Data analysis, Writing - original draft. **Legesse Hagos**: Data acquisition, Writing - review & editing. **Tamirat Kore**: Data acquisition, Writing - review & editing. **Shimelish A. Emiru**: Methodology, Writing - review & editing, supervision.

Acknowledgments

The authors wish to thank the Ethiopian Institute of Agricultural Research for providing the necessary facilities for carrying out the study.

Author's information

[1]Food Science and Nutrition Research, National Fishery and Aquatic Life Research Center, Ethiopian Institute of Agricultural Research, P. O. Box 64, Sebeta, Ethiopia; [2]Food Science and Nutrition Research, Debre Zeit Agricultural Research Center, Ethiopian Institute of Agricultural Research, Debre Zeit, Ethiopia; [3]Department of Food Engineering, School of Chemical and Bioengineering, Addis Ababa University, P.O. Box 33381, Addis Ababa, Ethiopia.

References

1. Berrueta LA, Alonso-Salces RM, Héberger K (2007) Supervised pattern recognition in food analysis. J Chromatogr A 1158(1–2):196–214
2. Brereton RG (2003) Chemometrics: Data analysis for the laboratory and chemical plant. John Wiley & Sons
3. Das S, Goswami B, Biswas K (2016) Milk Adulteration and Detection: A Review. Sensor Letters 14(1):4–18. https://doi.org/10.1166/sl.2016.3580
4. Dos Santos PM, Pereira-Filho ER (2013) Digital image analysis–an alternative tool for monitoring milk authenticity. Anal Methods 5(15):3669–3674
5. Faraz A, Lateef M, Mustafa MI, Akhtar P, Yaqoob M, Rehman S (2013) Detection of adulteration, chemical composition and hygienic status of milk supplied to various canteens of educational institutes and public places in Faisalabad. JAPS Journal of Animal Plant Sciences 23(1 Supplement):119–124
6. Handford CE, Campbell K, Elliott CT (2016) Impacts of Milk Fraud on Food Safety and Nutrition with Special Emphasis on Developing Countries. Comprehensive Reviews in Food Science Food Safety 15(1):130–142. https://doi.org/10.1111/1541-4337.12181
7. Kandpal SD, Srivastava AK, Negi KS (2012) ESTIMATION OF QUALITY OF RAW MILK (OPEN & BRANDED) BY MILK ADULTERATION TESTING KIT. Indian Journal of Community Health 24(3):188–192
8. Kobek JA (2017) Vision based model for identification of adulterants in milk. Strathmore University
9. Lopes JF, Ludwig L, Barbin DF, Grossmann MVE, Barbon S (2019) Computer vision classification of barley flour based on spatial pyramid partition ensemble. Sensors 19(13):2953
10. Massart DL, Vandeginste BG, Buydens LM, Lewi PJ, Smeyers-Verbeke J, Jong SD (1998) Handbook of chemometrics and qualimetrics. Elsevier Science Inc
11. Moore JC, Spink J, Lipp M (2012) Development and Application of a Database of Food Ingredient Fraud and Economically Motivated Adulteration from 1980 to 2010. J Food Sci 77(4):R118–R126. https://doi.org/10.1111/j.1750-3841.2012.02657.x
12. Nascimento CF, Santos PM, Pereira-Filho ER, Rocha FR (2017) Recent advances on determination of milk adulterants. Food Chem 221:1232–1244
13. Park YW, Haenlein GF, Ag DS (2013) Milk and dairy products in human nutrition. *Wilet-Blackwell. A John Wiley & Sons, Ltd., Publication*, *700*
14. Poonia A, Jha A, Sharma R, Singh HB, Rai AK, Sharma N (2017) Detection of adulteration in milk: A review. Int J Dairy Technol 70(1):23–42. https://doi.org/10.1111/1471-0307.12274
15. Shabir Barham G (2014) Detection and Extent of Extraneous Water and Adulteration in Milk Consumed at Hyderabad, Pakistan. Journal of Food Nutrition Sciences 2(2):47. https://doi.org/10.11648/j.jfns.20140202.15
16. Soomro AA, Khaskheli M, Memon MA, Barham GS, Haq IU, Fazlani SN, Khan IA, Lochi GM, Soomro RN (2014) Study on adulteration and composition of milk sold at Badin. Intl J Res Appl Nat Social Sci 2(9):57–70
17. Tang J, Alelyani S, Liu H (2014) Data classification: Algorithms and applications. *Data Mining and Knowledge Discovery Series, CRC Press (2014)*, 37–64

Figures

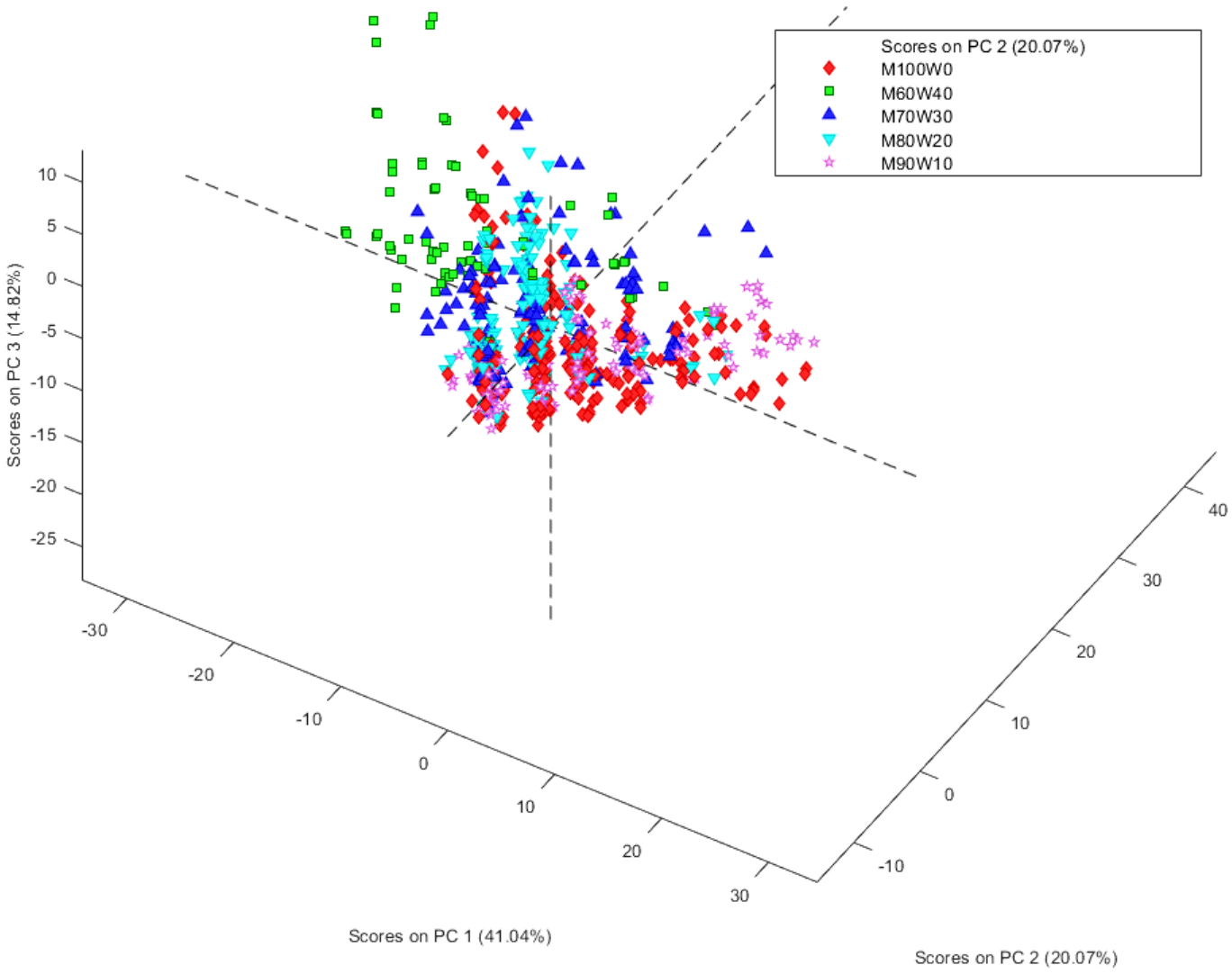

Figure 1

3-D score plot of adulterated milk samples