



**QUEEN'S
UNIVERSITY
BELFAST**

Innovative and rapid analysis for rice authenticity using hand-held NIR spectrometry and chemometrics

Teye, E., Amuah, C. L. Y., McGrath, T., & Elliott, C. (2019). Innovative and rapid analysis for rice authenticity using hand-held NIR spectrometry and chemometrics. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 217, 147-154. <https://doi.org/10.1016/j.saa.2019.03.085>

Published in:

Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy

Document Version:

Peer reviewed version

Queen's University Belfast - Research Portal:

[Link to publication record in Queen's University Belfast Research Portal](#)

Publisher rights

© 2019 Elsevier Ltd.

This manuscript version is made available under the CC-BY-NC-ND 4.0 license <http://creativecommons.org/licenses/by-nc-nd/4.0/>, which permits distribution and reproduction for non-commercial purposes, provided the author and source are cited.

General rights

Copyright for the publications made accessible via the Queen's University Belfast Research Portal is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy

The Research Portal is Queen's institutional repository that provides access to Queen's research output. Every effort has been made to ensure that content in the Research Portal does not infringe any person's rights, or applicable UK laws. If you discover content in the Research Portal that you believe breaches copyright or violates any law, please contact openaccess@qub.ac.uk.

Open Access

This research has been made openly available by Queen's academics and its Open Research team. We would love to hear how access to this research benefits you. – Share your feedback with us: <http://go.qub.ac.uk/oa-feedback>

1 **Innovative and rapid analysis for rice authenticity using hand-held NIR spectrometry and**
2 **chemometrics**

3 **Ernest Teye^{1&2}, Charles. L. Y. Amuah³, Terry McGrath² and Christopher Elliott²*

4 ¹University of Cape Coast, School of Agriculture, Department of Agricultural Engineering, Cape
5 Coast, Ghana

6 ²Institute for Global Food Security, Queen’s University Belfast, Northern Ireland, U.K.

7 ³University of Cape Coast, School of Physical Sciences, Department of Physics, Laser and Fibre
8 Optics Centre, Cape Coast, Ghana

9 Tel :+233-243170302/+233-206969565

10 *Email: teyernest@gmail.com / ernest.teye@ucc.edu.gh

11 **Abstract**

12 Rice is the second most important food staple worldwide and the demand will continue to
13 increase with the growth of the world population. As reports grow that frauds is prevalent in
14 many supply chains there is the need for an effective and rapid technique for monitoring the
15 authenticity and quality of rice. This study investigated the novel application of hand-held NIR
16 spectrometry coupled to chemometric for the estimation of rice authenticity and quality in real
17 time. A total of 520 Rice samples from different quality grades (high quality, mid quality and
18 low quality) and different countries (Ghana, Thailand, and Vietnam) of origin were used. Among
19 the pre-processing methods used multiplicative scatter correction (MSC) was found to be
20 superior. Principal component analysis (PCA) was used to extract relevant information from the
21 spectral data set and the results showed that rice samples of different categories could be clearly
22 clustered under the first three PCs using the MSC preprocessing method. The performance of K-
23 nearest neighbor (KNN) revealed that for authentication of rice quality grades, the classification
24 rate gave 91.62% and 91.81% in training set and prediction set respectively while identification
25 rate based on different country of origin was 90.84% and 90.64% in both training set and
26 prediction set respectively. For the differentiation of local rice from the imported, KNN and
27 SVM all had 100% in both the training set and prediction set. These gives very strong evidence
28 that hand-held spectrometry coupled with MSC-PCA-KNN could successfully be used to

29 provide rapid and nondestructive classification of rice samples according to different quality
30 grades, geographical origin and imported versus locally produced rice. This technique could
31 enhance the work of quality control inspectors both from industry and regulatory perspectives for
32 the rapid detection of rice integrity and fraud issues.

33 **Keywords:** Rice, Hand-held NIR spectroscopy, Chemometrics, Quality, Authenticity

34 **1.0 INTRODUCTION**

35 Rice (*Oryza sativa* L.) has increasingly become a hugely important staple food worldwide and its
36 consumption has surged as a result of population growth, changing food preferences and
37 urbanization [1]. Globally the top five leading net exporters of rice are Thailand, Vietnam, India,
38 China, and Pakistan. Generally, Asian countries produce the largest quantities of rice worldwide
39 while in Latin America, Middle East and African countries such as Ghana, Nigeria, Ivory Coast,
40 has shown considerable increase in the importation and consumption of rice. Rice is consumed
41 by more than 3.5 billion people (almost half of the world's population) and in 2015 West Africa
42 countries imported more than 7.6 million tonnes of rice amounting to USD 4 billion [1]. Thus,
43 Africa countries are net rice importers and given the increasing rice consumption trends in Africa
44 this will continue to grow [2]. Due to the large diversity of rice in terms of types and country of
45 origin and the increasing volumes being imported, it is extremely difficult to identify the quality
46 and authenticity status at the ports and in the markets and has led to growing rumours that
47 substandard product is flooding the African market places.

48 What is clear is that, some unscrupulous actors in the rice supply chain are selling sub-standard
49 rice to consumers and very often these unscrupulous activities go undetected, because the
50 existing analytical methods are cumbersome, time consuming, expensive and can involve
51 substantial use of chemicals. For instance the DNA finger-printing method based on the genetic
52 characteristic of rice has shown to be inappropriate for on-site application, it also involves labour
53 intensive and skilled expertise [3, 4]. The procedure is also quite cumbersome and laboratory
54 based and limits the number of sample that can be tested due to the cost involved. Other methods
55 which are used are sensory examination by well-trained inspector [5, 6], enzyme linked
56 immunosorbent assays, qualitative polymerase chain reaction, high performance liquid
57 chromatography and gas chromatography [7, 8]. However all these methods has many

58 shortcomings, such as high cost, difficulty in use, and substantial time gap between sampling and
59 results generation [9].

60 In a competitive market however, fast, accurate evaluation and classification of rice quality
61 would lead to better decision making in terms of better quality control management. This would
62 further give consumers a stronger assurance around product quality. NIR spectroscopy could
63 provide this technology for such rapid examination. NIR spectroscopy is an excellent, rapid
64 analytical method that has been used in several industries such as pharmaceutical, petrochemical,
65 agricultural, and food processing. This method coupled with chemometrics has far-fetching
66 advantages over the traditional analytical methods; it is rapid, nondestructive, reliable and
67 involves no chemical usage [10, 11]. Also it requires little or no sample preparation and can be
68 used in an on-line monitoring tool. Previous research has demonstrated that NIR spectroscopy
69 has high potential for nondestructive measurement of qualitative and quantitative attributes in
70 different agricultural and food products [12, 13].

71 For rice quality evaluation, NIR spectroscopy has been used for; discrimination of rice [9, 14],
72 rice quality measurement [15-18] and classification of cultivars [19], authenticity detection [20],
73 prediction of protein and amylase content [17, 21], detection of wax rice [22] and prediction of
74 eating quality [23]. However, with the recent development of miniaturized or portable NIR
75 spectrometers, there are no studies to date on the use of portable NIR devices for rice quality
76 control. However a few studies have investigated the technology in terms of its use for food
77 product quality measurement such as: identification of barley, chickpea and sorghum cultivar
78 [24], determination of quality parameters in fruits; mango & pineapple [11, 25], determination of
79 sucrose in sugar beet [26] and identification of coffee [27]. However, no studies until that
80 reported herein are available for the use of hand-held NIR spectroscopy for rapid and
81 nondestructive examination of rice quality and authenticity. Furthermore, there is no discussion
82 on the use of different preprocessing techniques and multivariate classification methods for
83 accurate determination of rice quality and authenticity. With the current incidences of rice fraud
84 being reported worldwide such as in the media about ‘plastic rice’ and ‘toxic rice’ etc, this
85 research could provide the urgently required reliable analytical technique for rapid detection of
86 rice quality and authenticity. It would be particularly helpful in developing countries where the
87 majority of these regions suffer from poor laboratory infrastructure and inadequate technical

88 know-how. The current research therefore seeks to develop a very reliable, nondestructive
89 prediction model using a hand-held NIR spectrometer coupled with chemometrics for rapid and
90 nondestructive evaluation of rice quality and authenticity. Also, there is the potential of this
91 model to be imported into smart phone base for on-site applications.

92 **2.0 MATERIALS AND METHODS**

93 **2.1 Sample preparation**

94 In this study, rice samples were collected from local millers, and recognized retailers around
95 seven regions in Ghana, while the other samples (imports from Thailand and Vietnam) were
96 bought from recognised super-markets in Ghana. In all a total of 520 samples of rice grains were
97 used in the study. For quality grades: 161 high quality, 250 mid quality and 109 low quality. For
98 country of origin: 329 Ghana, 99 Thailand, 92 Vietnam, while the others were 191 imported
99 samples and 329 local samples. These samples were bagged in sealable polythene bags and
100 were grouped into three quality categories namely; high quality, mid quality and low quality.
101 These quality groupings were based on the cost price, and confirmed by sensory analysis using a
102 skilled panelist to exam the physical appearance, cleanliness, colour, and aroma on a six (1-6)
103 corresponding level as done by other authors [28]. All samples were (three replication each)
104 were sent to the laboratory for further analysis.

105 **2.2 Sample spectral acquisition**

106 The spectrum of each rice sample was collected using a hand-held spectrometer (SCIO™) with
107 spectral range between 740 nm and 1070 nm in a 1-nm resolution for spectral data recording.
108 Samples (50 g) were collected into glass containers and scanned three times after rotating the
109 cup. The whole process was carried out at ambient temperature. The set-up of the scanning
110 procedure is shown in the graphical abstract.

111 **2.3 Software device**

112 Spectral data recordings stored in a cloud-based dataset with their corresponding reference value
113 for time of scanning were downloaded using a research license acquired from SCIO lab and
114 imported to Matlab version 9.5.0 (Mathworks Inc., USA) with windows 10 Basic for data
115 processing (All preprocessing treatments and multivariate algorithms).

116 **2.4 Spectral preprocessing techniques**

117 The raw spectra profile from the rice samples are shown in Fig. 1.0 (A) and the mean spectra of
118 the various categories of the rice samples of interest are detailed in Fig. 1.0 (B to C). This
119 outlines the three categories of rice samples used in the study. The average of the triplicate scans
120 (whole spectral data set) were preprocessed before further analysis. The activity of preprocessing
121 the spectra data is an integral part of modelling to eliminate background information and noise
122 from the useful properties of the scanned samples [29, 30]. In this study, four spectra
123 preprocessing techniques were applied comparatively, namely: de-trending (DT), mean centering
124 (MC), multiplicative scatter correction (MSC), and standard normal variant (SNV) were
125 employed as the spectral models developed using unprocessed did not yield good results. These
126 few preprocessing techniques were selected after initial background studies. These preprocessing
127 techniques were undertaken so that the results from the model were based on the chemical
128 fingerprint from the spectral information acquired.

129 **2.5 Theory of chemometric techniques**

130 **2.5.1 De-trend (DT)**

131 DT was trailed because it is frequently applied to spectra to remove baseline drift and
132 curvilinearity for densely packed solids [31]. In detrending, a trend line is made from spectral
133 data set by least squares fitting and then the trend line is subtracted from the original spectrum to
134 correct for any defect.

135 **2.5.2 Mean centering (MC)**

136 MC performs calculation by averaging the spectrum of the data set and subtracting the average
137 from each spectrum [31].

138 **2.5.3 Multiplicative scatter correction (MSC)**

139 MSC is a well known technique used since 1983 for the removal of undesirable scatter effect
140 from spectra data matrix before modelling [29, 32]. It is also used to compensate the effect of
141 non-uniform scattering induced by diverse particle size, uneven distribution and other effects in
142 the spectral data and it works by linearizing each spectrum to an “ideal” spectrum, which

143 corresponds to the average spectrum of the calibration set and the average spectrum is fitted
144 through the method of least squares [33].

145 **2.5.4 Standard normal variant (SNV)**

146 SNV is similar to MSC, it is a transformational spectral treatment method normally applied to
147 remove the multiplicative interferences of scatter, particle size and the change of light distance
148 [34]. In SNV each individual spectrum is normalized to zero mean and unit variance [13, 34].

149 **2.5.5 Principal component analysis (PCA)**

150 PCA is an unsupervised pattern recognition method that; employs the technique of extracting
151 information from correlation matrices to visualize data trends in a dimensional scatter plot. The
152 unsupervised classification terminology means that; the samples are classified with no prior
153 knowledge, except the sensor signals. PCA expresses information contained in a dataset by
154 principal components or reduces the dimension of the data matrix and compresses the
155 information into interpretable variables called principal components (PCs) which are orthogonal
156 [35, 36]. PCA has been used to observe possible groupings in several sensor data set. The best
157 performing PCs normally show the most important information. Therefore, similar samples are
158 grouped closer to each other and vice versa. Graphical profile of PCA results normally provides
159 an initial output for determining the possible differences and similarities in a data set. PCA can
160 be used to identify combinations of variables that have the largest contribution to variations in
161 the data set, as these variables are retained in the first two or three PCs [36].

162 **2.6 Data partition**

163 The spectral data-set (520 samples) was downloaded and preprocessed with suitable techniques.
164 Furthermore, these data were divided into two subsets called: calibration set (347 samples); was
165 used to develop the model and prediction set (173 samples); was used for evaluating the actual
166 predictive ability of the constructed models. The members in each set were selected in order to
167 come to a 2/1 division of calibration set/prediction set. To avoid bias, members of the subset
168 were selected as follows: for every three samples, two spectra were randomly selected as the
169 calibration set while the remaining sample was the prediction set.

170 **2.7 Multivariate classification models**

171 The advancement in computers and electronics have contributed to making multivariate
172 calibration a very powerful tool for processing NIR spectral data as it overcomes the difficulty of
173 multi-collinearity and gives scientific statistical inferences for meaningful conclusions to
174 experimental results [37, 38]. However, choosing the best method can be quite a cumbersome
175 process as many are in existence. In this study; linear and non-linear algorithm were employed
176 such as; K-nearest neighbor (KNN) and Support vector machine (SVM) respectively.

177 **2.7.1 K-nearest neighbour (KNN)**

178 KNN is a linear and non-parametric tool [39] which works based on a distance function that
179 measures the difference or similarity between two stances [40]. KNN classifies an unknown
180 sample of the validation set according to the class which belongs to the majority of its K nearest
181 neighbour in the training set. In KNN, parameter K has a great impact on the classification
182 model; hence the choice of K is normally optimized by calculating the prediction potential with
183 several K values, preferably an odd number of small K values (3 or 5). Also to improve the
184 performance of KNN model, there is the need to simultaneously optimize PCs and parameter K
185 to derive a very good model. Furthermore, KNN technique presents some advantages such as it
186 is: free from statistical assumption, mathematically simple but achieving classification results as
187 good as other more complex pattern recognition method, and its effectiveness does not depend
188 on the space distribution of the classes. However, it is known that KNN cannot work well if large
189 differences are present in the number of samples in each class [41]. This therefore makes KNN
190 tool a suitable technique for modeling similar class groupings. Other authors used KNN for
191 modeling identification of tea grade groups [42] and discrimination of roasted tea [43].

192 **2.7.2 Support vector machine (SVM)**

193 SVM is supervised statistical learning theory applicable for both classification and regression
194 problems or ranking functions [44]. It has shown good performance for classifying high-
195 dimensional data when a limited number of training samples are available [45]. SVM processes
196 sensor data by obtaining the optimal boundary of two groups in a vector space independent on
197 the probabilistic arrangements of vectors in the training set. When the linear boundary in the low
198 dimension input space is not enough to separate the two classes, SVM can create a hyperplane
199 that allows linear separation in the higher dimension feature space. However, if the classes are
200 separated by non-linear boundary, then the kernel function is used to find the boundary by

201 mapping the non-separable data into a higher dimensional space and causes the classes to
202 become linearly separated. The strength of SVM over the others is that; it can achieve higher
203 generalization by maximizing the margin and it can support an efficient learning of non-linear
204 functions using the kernel trick [44]. Among the three kernel functions (sigmoid kernel,
205 polynomial kernel and Gaussian kernel), Gaussian kernel function is mostly employed because
206 of its simplicity and speed during its computation [41]. SVM is likely to have a better
207 generalization, thus can accurately classify testing data points. For more information refer [44].

208 **2.8 Model evaluation procedure**

209 In this study, all the multivariate classification models were tested using the performance
210 measure as the co-efficient of determination in a validation set for discrimination rate (%)

211 **3.0 RESULTS AND DISCUSSION**

212 **3.1 Spectral presentation**

213 All the rice samples used in this study were scanned three times from different angles after
214 rotating the sampling cup 45⁰ clockwise. The fingerprint from the spectra was used to create the
215 statistical models. Figure 1 shows A= raw spectra, B= average spectra for quality categories, C=
216 average spectra for country of origin and D= imported versus local rice samples. As seen from
217 this figure, the spectra share very similar absorbance patterns in the 740 – 1070 nm range. It
218 appears they could not be easily distinguished by interpretation of the raw data. However, these
219 spectra did contain very useful yet unexploded information, therefore the need to pretreat the
220 spectra data with suitable pretreatment techniques was apparent. The selection of the most
221 suitable pretreatment techniques can be quite cumbersome, and this was mainly performed using
222 a trial and error approach involving several selections [29, 46]. Finally, three preprocessing
223 techniques were selected and were compared with unprocessed (raw) spectral profile.

224 **Figure 1.0** NIR spectra of rice samples; (A) = raw, (B) = mean of different quality categories,
225 (C) = country of origin and (D) = imported versus local

226 **3.2 Principal component analysis (PCA)**

227 The PCA results from the three attributes studied; quality grades, countries of origin and
228 imported versus locally grown revealed that the topmost first three PCs gave clear separations in
229 all the attributes studied with MSC providing the best preprocessing results in PCA among the

230 three others methods (Raw, MC, SNV). For MSC-PCA method, all the samples clustered well
231 along the first three PCs plane.

232 **3.3 Classification by quality grades**

233 The score plots obtained after MSC+PCA with all the spectral fingerprints based on the quality
234 grades are shown in Figure 2. The first two principal components (PCs); PC1 and PC2 can
235 explain 86.78% and 7.63% of the variance respectively giving a total accumulation contribution
236 of 94.41% total variance for the 520 samples used in this study. A neat clustering of the quality
237 grades were observed for MSC+PCA while a faint clustering was observed for MC, De-trending
238 and SNV. This phenomenon could be explained as MSC has the ability to correct scattering
239 effect. KNN and SVM were applied separately after PCA to perform the identification between
240 the three quality grades. The results of the identification rate are summarized in Table 1. From
241 this table, it can be seen that the derived model; MSC+PCA+KNN gave an optimal identification
242 rate above 90% in the training set and prediction set as compared to the others.

243 **Figure 2.0** PCA score plot of rice sample of different quality preprocessed (A) MC, (B) DT, (C)
244 SNV and (D) MSC

245 **Table 1.0** Performance of multivariate classification models based on quality grades

246 **3.4 Classification by country of origin**

247 After the identification of the quality grades, the possibility to discriminate between rice from
248 different countries was investigated. This is particularly important because some consumers
249 prefer rice from other countries over others. The variance in the data was mainly due to the
250 country of origin (Ghana, Thailand, and Vietnam). For the PCA score plot shown in Figure 3, the
251 best in-term of neat clustering was MSC-PCA where PC1, PC2 and PC3 can explain 80.97%,
252 14.15% and 2.09% of the variance respectively. With a total accumulation contribution of
253 97.22% variance for the 520 samples used. KNN and SVM were applied on all the samples to
254 perform the discrimination of countries of origin. The results of the discrimination are
255 summarized in Table 2. From this table it can be seen that model MSC+PCA+KNN gave the best
256 identification rate of 90.84% and 90.64% in both the training set and prediction set as compare to
257 the other models.

258 **Figure 3.0** PCA score plot of rice sample from different origin preprocessed (A) MC, (B) DT,
259 (C) SNV and (D) MSC

260 **Table 2.0** Performance of multivariate classification models based on country of origin

261 **3.5 Classification of imported versus local**

262 After the classification of quality grades and country of origin, the ability to differentiate
263 between imported rice and local produce was investigated. This is particularly useful with the
264 current incidences of imported rice fraud often reported in the media. Also the total ban on rice
265 importation by some African countries and the encouragement of local production to reduce
266 imports calls for easy and inexpensive testing techniques. Furthermore, cheap imports are
267 competing with high quality local brands. Hence the potential of food fraudsters operating to
268 gain economic advantage is large. For imported versus local (as seen from Figure 4) PC1, PC2
269 and PC3 gave 98.89%, 0.84% and 0.11% of the variance respectively, giving a total
270 accumulation contribution of 99.84% variance for the 520 samples used. Again MSC+PCA gave
271 a visible cluster trend in the PCA score plot of imported rice versus local rice. The derived
272 model; MSC+PCA+KNN and MSC+PCA+SVM for differentiating imported from local gave
273 100% differentiation rate in both the training set and prediction set at 2 PCs respectively (as seen
274 from Table 3) compared to the others. The results proved that KNN and SVM could be an
275 excellent model for the classification of rice samples from different origins.

276 **Figure 4.0** PCA score plot of rice sample local & imported preprocessed (A) MC, (B) DT, (C)
277 SNV and (D) MSC

278 **Table 3.0** Performance of multivariate classification models based on imported versus local

279 **4.0 General discussion**

280 The NIR scanning process on the rice produced spectra that showed multiple bands and a few
281 peaks as seen from the spectral profile presented. These bands are made up of overtones and
282 combinations of fundamental vibrations which correspond to organic properties that provide
283 unique characteristics or a fingerprint of the rice samples used. To the naked eye they appear
284 very similar, though there exists many differences with useful and non-useful information. To
285 extract the useful and unique information, chemometric techniques were applied. Firstly several

286 (MC, SNV, DT, MSC) preprocessing treatments were performed and compared prior to PCA.
287 PCA as an unsupervised mathematical technique was used for extracting information from
288 correlation matrices [47]. This gave visible cluster trends to show the strength of each
289 preprocessing technique as shown in Figure 2, 3 & 4. MSC was found to give neat cluster trends
290 and hence superior to the others applied in this study. This could be explained in that, MSC has
291 the power to remove undesirable scatter effect of spectra from a data matrix before modelling
292 [29] and this has proved highly useful in this study. Also, the scatter effect is known to influence
293 modelling as it contains unwanted information for prediction. The groupings found in the PCA
294 cluster plot could then be explained further by the chemical properties in each as a result of the
295 differences in quality grade, geographical origin, pre-harvest and postharvest practices [31]. The
296 contribution of the best performing three PCs were 98.67%, 97.22% and 99.84% for quality
297 grade, country of origin and imported versus local respectively. However, PCA is not a
298 classification tool but only functions by reducing dimensionality while preserving as much
299 variance in a high dimensional space. Thus making it possible to extract useful information from
300 high-dimensional data since the spectral data forms the array of correlated variables which
301 contains overlapped information [47]. Hence, the preprocessed data (MSC-PCA data; which
302 provided useful unique information) were further developed by a supervised pattern recognition
303 technique known as KNN and SVM. The classification rates for the KNN models were above
304 90% in both the training set and prediction set respectively for all the types of rice studied. This
305 could be explained by the fact that, KNN; as a simple linear and non-parametric tool classified
306 the groups according to the classes which belongs to the majority of its K nearest neighbor. In
307 this study PCs were also used as an input data in KNN and this enhanced the efficiencies of the
308 KNN models. This was because factors (PCs & K) were simultaneously optimized by cross-
309 validation method [43] as this act is useful for deriving a very good KNN model [31]. Also the
310 number of PCs used in the KNN models suggests that it is simpler and normally, lower values of
311 parameter K (3 or 5) and PCs are preferred as higher number of PCs included in training a model
312 brings out too much redundant information which inescapably influences the robustness of the
313 model [38, 48]. KNN is also known to perform equally to (or even better than) other more
314 complex pattern recognition techniques but its effectiveness does not depend on the space
315 distribution of the classes [41]. The performance of KNN model in this study was similar to the
316 findings by other authors; classification of tea quality grades [42] discrimination of roasted tea

317 [43]. However, KNN cannot work well if large differences are present in the spectral dataset as it
318 is known to be a ‘lazy’ training algorithm [41]. SVM was also employed in this study. From the
319 results, SVM model also showed comparatively good results as KNN with 100% classification
320 rate in both the training set and prediction set for only classifying imported versus local as shown
321 in Table 3. This optimal result could be explained in that, SVM model embodies structural risk
322 minimization principle where upper boundary is reduced on the expected risk and also possess
323 the power of a better generalization, thus classifying accurately in the prediction set [43, 44, 49].
324 Comparing the two models (KNN and SVM) it must be emphasized that since KNN being a lazy
325 learning method and the class probability estimation is based on a simple voting of a “good
326 value” for k, as well as stores all the training data, the prediction stage is often slow hence
327 require high memory and quite expensive [40, 50, 51]. However in this study it proved very
328 useful for all the classification problems.

329 **5.0 CONCLUSION**

330 This study has revealed, for the first time, that hand-held NIR spectrometer coupled with the
331 appropriate chemometrics could be used for rapid and nondestructive detection of rice
332 authenticity and quality. The systematic selection of different preprocessing methods (MC, DT,
333 SNV, and MSC) with PCA and modeling with KNN and SVM multivariate calibration model
334 showed that MSC+PCA plus KNN showed superiority in this study with more than 90%
335 classification rate for all categories of rice samples studied. Generally, it could be concluded that
336 hand-held spectrometer together with appropriate multivariate calibration model could be
337 exploited for quick and nondestructive detection of rice quality and authenticity. Furthermore,
338 there is a potential for this model to be imported into smart phone for effective quality control
339 measurements in the rice industry and by regulators as compared with the time-consuming wet
340 chemistry analytical methods. Further studies are required for the prediction of other quality
341 attributes in rice.

342 **Acknowledgement**

343 The authors highly appreciate the financial support provided by Agilent foundation (sub-contract
344 No. 4368). Proof reading assistance provided by Mrs. Winifred Akpene Teye is also
345 acknowledged. The kind contribution given by University of Cape Coast is fully appreciated.

346 **References**

- 347 1. Fiamohe, R., et al., *How Can West African Rice Compete in Urban Markets? A Demand*
348 *Perspective for Policymakers*. EuroChoices, 2018.
- 349 2. Kormawa, P., et al. *Global Rice Trade: Dynamics, Policy Conflicts and Strategies in Africa*. in
350 *Conference on International Agricultural Research for Development*. Herausgegeben von Africa
351 *Rice Center (WARDA)*. Cotonou. 2005.
- 352 3. Shanthi, P., et al., *DNA finger printing of salt tolerant and susceptible genotypes using*
353 *microsatellite markers in rice (Oryza sativa L.)*. Int J Plant Breed Genet, 2012. **6**: p. 206-16.
- 354 4. Jeung, J., et al., *Fingerprinting temperate japonica and tropical indica rice genotypes by*
355 *comparative analysis of DNA markers*. Euphytica, 2005. **146**(3): p. 239-251.
- 356 5. Kwak, H.S., et al., *Identification of key sensory attributes for consumer acceptance and*
357 *instrumental quality of aseptic - packaged cooked rice*. International Journal of Food Science &
358 *Technology*, 2015. **50**(3): p. 691-699.
- 359 6. Champagne, E.T., et al., *Important sensory properties differentiating premium rice varieties*. Rice,
360 2010. **3**(4): p. 270-281.
- 361 7. Wu, J., et al., *Metabolite profiles of rice cultivars containing bacterial blight-resistant genes are*
362 *distinctive from susceptible rice*. Acta Biochim Biophys Sin, 2012. **44**(8): p. 650-659.
- 363 8. Kemnitz, D., S. Kolb, and R. Conrad, *Phenotypic characterization of Rice Cluster III archaea*
364 *without prior isolation by applying quantitative polymerase chain reaction to an enrichment*
365 *culture*. Environmental Microbiology, 2005. **7**(4): p. 553-565.
- 366 9. Long, Z., et al., *Discrimination of transgenic rice based on near infrared reflectance spectroscopy*
367 *and partial least squares regression discriminant analysis*. Rice science, 2015. **22**(5): p. 245-249.
- 368 10. Cen, H., Y. He, and M. Huang, *Combination and comparison of multivariate analysis for the*
369 *identification of orange varieties using visible and near infrared reflectance spectroscopy*.
370 *European Food Research and Technology*, 2007. **225**(5-6): p. 699-705.
- 371 11. Marques, E.J.N., et al., *Rapid and non-destructive determination of quality parameters in the*
372 *'Tommy Atkins' mango using a novel handheld near infrared spectrometer*. Food chemistry,
373 2016. **197**: p. 1207-1214.
- 374 12. Huang, X., et al., *Rapid measurement of total polyphenols content in cocoa beans by data fusion*
375 *of NIR spectroscopy and electronic tongue*. Analytical Methods, 2014. **6**(14): p. 5008-5015.
- 376 13. Nicolai, B.M., et al., *Nondestructive measurement of fruit and vegetable quality by means of NIR*
377 *spectroscopy: A review*. Postharvest biology and technology, 2007. **46**(2): p. 99-118.
- 378 14. Chen, K. and M. Huang, *Prediction of milled rice grades using Fourier transform near-infrared*
379 *spectroscopy and artificial neural networks*. Journal of cereal science, 2010. **52**(2): p. 221-226.
- 380 15. Chang, R., et al. *Research of rice-quality based on computer vision and near infrared*
381 *spectroscopy*. in *International Conference on Computer and Computing Technologies in*
382 *Agriculture*. 2009. Springer.
- 383 16. Caporaso, N., M.B. Whitworth, and I.D. Fisk, *Near-Infrared spectroscopy and hyperspectral*
384 *imaging for non-destructive quality assessment of cereal grains*. Applied Spectroscopy Reviews,
385 2018: p. 1-21.
- 386 17. Zhang, B., et al., *Prediction of the amino acid composition in brown rice using different sample*
387 *status by near-infrared reflectance spectroscopy*. Food Chemistry, 2011. **127**(1): p. 275-281.
- 388 18. Natsuga, M. and S. Kawamura, *Visible and near-infrared reflectance spectroscopy for*
389 *determining physicochemical properties of rice*. Transactions of the ASABE, 2006. **49**(4): p. 1069-
390 1076.
- 391 19. Kong, W., et al., *Rice seed cultivar identification using near-infrared hyperspectral imaging and*
392 *multivariate data analysis*. sensors, 2013. **13**(7): p. 8916-8927.
- 393 20. Chen, H., C. Tan, and Z. Lin, *Authenticity Detection of Black Rice by Near-Infrared Spectroscopy*
394 *and Support Vector Data Description*. International journal of analytical chemistry, 2018. **2018**.

- 395 21. Xie, L., et al., *Optimisation of near-infrared reflectance model in measuring protein and amylose*
396 *content of rice flour*. Food chemistry, 2014. **142**: p. 92-100.
- 397 22. Li, B., et al., *Detection of waxed rice using Visiblener infrared hyperspectral imaging*. J. Food
398 Nutr. Res, 2016. **4**: p. 267-275.
- 399 23. Siriphollakul, P., et al., *Eating quality evaluation of Khao Dawk Mali 105 rice using near-infrared*
400 *spectroscopy*. LWT-Food Science and Technology, 2017. **79**: p. 70-77.
- 401 24. Kosmowski, F. and T. Worku, *Evaluation of a miniaturized NIR spectrometer for cultivar*
402 *identification: The case of barley, chickpea and sorghum in Ethiopia*. PloS one, 2018. **13**(3): p.
403 e0193620.
- 404 25. Chia, K.S., H.A. Rahim, and R.A. Rahim, *Prediction of soluble solids content of pineapple via non-*
405 *invasive low cost visible and shortwave near infrared spectroscopy and artificial neural network*.
406 Biosystems Engineering, 2012. **113**(2): p. 158-165.
- 407 26. Pan, L., et al., *Determination of sucrose content in sugar beet by portable visible and near-*
408 *infrared spectroscopy*. Food chemistry, 2015. **167**: p. 264-271.
- 409 27. Correia, R.M., et al., *Portable near infrared spectroscopy applied to quality control of Brazilian*
410 *coffee*. Talanta, 2018. **176**: p. 59-68.
- 411 28. Lu, L., et al., *Determination of rice sensory quality with similarity analysis-artificial neural*
412 *network method in electronic tongue system*. RSC Advances, 2015. **5**(59): p. 47900-47908.
- 413 29. Coronel-Reyes, J., et al., *Determination of egg storage time at room temperature using a low-*
414 *cost NIR spectrometer and machine learning techniques*. Computers and Electronics in
415 Agriculture, 2018. **145**: p. 1-10.
- 416 30. Blanco, M. and I. Villarroya, *NIR spectroscopy: a rapid-response analytical tool*. TrAC Trends in
417 Analytical Chemistry, 2002. **21**(4): p. 240-250.
- 418 31. Teye, E., et al., *Rapid differentiation of Ghana cocoa beans by FT-NIR spectroscopy coupled with*
419 *multivariate classification*. Spectrochimica Acta Part A: Molecular and Biomolecular
420 Spectroscopy, 2013. **114**: p. 183-189.
- 421 32. Martens, H., S. Jensen, and P. Geladi. *Multivariate linearity transformation for near-infrared*
422 *reflectance spectrometry*. in *Proceedings of the Nordic symposium on applied statistics*. 1983.
423 Stokkand Forlag Publishers Stavanger, Norway.
- 424 33. Wang, H., et al., *Fruit quality evaluation using spectroscopy technology: a review*. Sensors, 2015.
425 **15**(5): p. 11889-11927.
- 426 34. Barnes, R., M.S. Dhanoa, and S.J. Lister, *Standard normal variate transformation and de-trending*
427 *of near-infrared diffuse reflectance spectra*. Applied spectroscopy, 1989. **43**(5): p. 772-777.
- 428 35. Hong, X. and J. Wang, *Detection of adulteration in cherry tomato juices based on electronic nose*
429 *and tongue: Comparison of different data fusion approaches*. Journal of Food Engineering, 2014.
430 **126**: p. 89-97.
- 431 36. Lavine, B.K. and N. Mirjankar, *Clustering and classification of analytical data*. Encyclopedia of
432 Analytical Chemistry: Applications, Theory and Instrumentation, 2006.
- 433 37. Roggo, Y., et al., *A review of near infrared spectroscopy and chemometrics in pharmaceutical*
434 *technologies*. Journal of pharmaceutical and biomedical analysis, 2007. **44**(3): p. 683-700.
- 435 38. Teye, E., et al., *Estimating cocoa bean parameters by FT-NIRS and chemometrics analysis*. Food
436 chemistry, 2015. **176**: p. 403-410.
- 437 39. Duda, R.O., P.E. Hart, and D.G. Stork, *Pattern classification*. 2012: John Wiley & Sons.
- 438 40. Jiang, L., et al. *Survey of improving k-nearest-neighbor for classification*. in *Fuzzy Systems and*
439 *Knowledge Discovery, 2007. FSKD 2007. Fourth International Conference on*. 2007. IEEE.
- 440 41. Berrueta, L.A., R.M. Alonso-Salces, and K. Héberger, *Supervised pattern recognition in food*
441 *analysis*. Journal of chromatography A, 2007. **1158**(1-2): p. 196-214.

- 442 42. Chen, Q., J. Zhao, and S. Vittayapadung, *Identification of the green tea grade level using*
443 *electronic tongue and pattern recognition*. Food Research International, 2008. **41**(5): p. 500-504.
- 444 43. Chen, Q., J. Zhao, and H. Lin, *Study on discrimination of Roast green tea (Camellia sinensis L.)*
445 *according to geographical origin by FT-NIR spectroscopy and supervised pattern recognition*.
446 Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 2009. **72**(4): p. 845-850.
- 447 44. Yu, H. and S. Kim, *SVM tutorial—classification, regression and ranking*, in *Handbook of Natural*
448 *computing*. 2012, Springer. p. 479-506.
- 449 45. Tarabalka, Y., et al., *SVM and MRF-based method for accurate classification of hyperspectral*
450 *images*. IEEE Geoscience and Remote Sensing Letters, 2010. **7**(4): p. 736-740.
- 451 46. Rinnan, Å., F. van den Berg, and S.B. Engelsen, *Review of the most common pre-processing*
452 *techniques for near-infrared spectra*. TrAC Trends in Analytical Chemistry, 2009. **28**(10): p. 1201-
453 1222.
- 454 47. Mehdizadeh, S.A., et al., *An intelligent system for egg quality classification based on visible-*
455 *infrared transmittance spectroscopy*. Information Processing in Agriculture, 2014. **1**(2): p. 105-
456 114.
- 457 48. Chen, Q., et al., *Nondestructive identification of tea (Camellia sinensis L.) varieties using FT-NIR*
458 *spectroscopy and pattern recognition*. Czech J Food Sci, 2008. **26**: p. 360-367.
- 459 49. Lin, H., et al., *Determination of free amino acid content in Radix Pseudostellariae using near*
460 *infrared (NIR) spectroscopy and different multivariate calibrations*. Journal of pharmaceutical
461 and biomedical analysis, 2009. **50**(5): p. 803-808.
- 462 50. Thirumuruganathan, S., *A detailed introduction to K-nearest neighbor (KNN) algorithm*.
463 Retrieved on July, 2010. **21**: p. 2015.
- 464 51. Guo, G., et al. *KNN model-based approach in classification*. in *OTM Confederated International*
465 *Conferences" On the Move to Meaningful Internet Systems"*. 2003. Springer.

466

467

468 **Table 1.0** Performance of multivariate classification models based on quality grades

Model		Number of principal components	Correct classification rate (%)	
			Training set (347)	Prediction set (173)
KNN	MC	7	69.59	76.02
	MSC	7	91.62	91.81
	DT	6	55.36	56.72
	SNV	5	77.39	78.95
SVM	MC	6	64.33	64.33
	MSC	5	87.52	86.55
	DT	5	48.34	47.37
	SNV	7	64.33	63.74

469

470 **Table 2.0** Performance of multivariate classification models based on country of origin

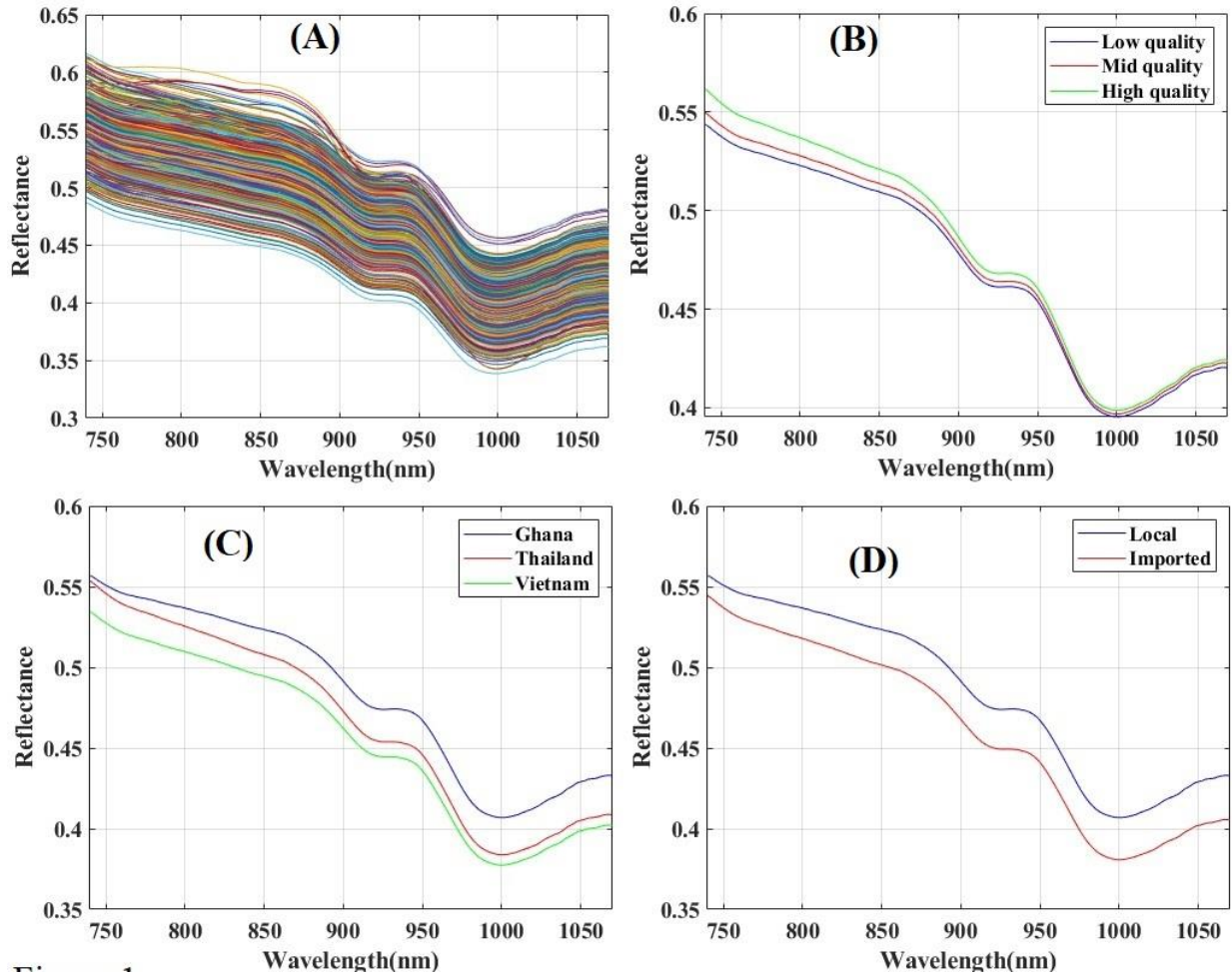
Model		Number of principal components	Correct classification rate (%)	
			Training set (347)	Prediction set (173)
KNN	MC	7	75.63	72.51
	MSC	6	90.84	90.64
	DT	7	52.05	56.73
	SNV	5	86.35	87.36
SVM	MC	7	79.14	78.36
	MSC	6	82.46	83.63
	DT	5	64.91	64.91
	SNV	7	79.92	80.70

471

472 **Table 3.0** Performance of multivariate classification models based on imported versus local

Model		Number of principal components	Correct classification rate (%)	
			Training set (347)	Prediction set (173)
KNN	MC	5	90.84	91.81
	MSC	2	100.00	100.00
	DT	5	64.72	66.08
	SNV	5	96.30	96.49
SVM	MC	5	95.32	95.32
	MSC	2	100.00	100.00
	DT	5	62.77	60.23
	SNV	7	95.91	97.08

473



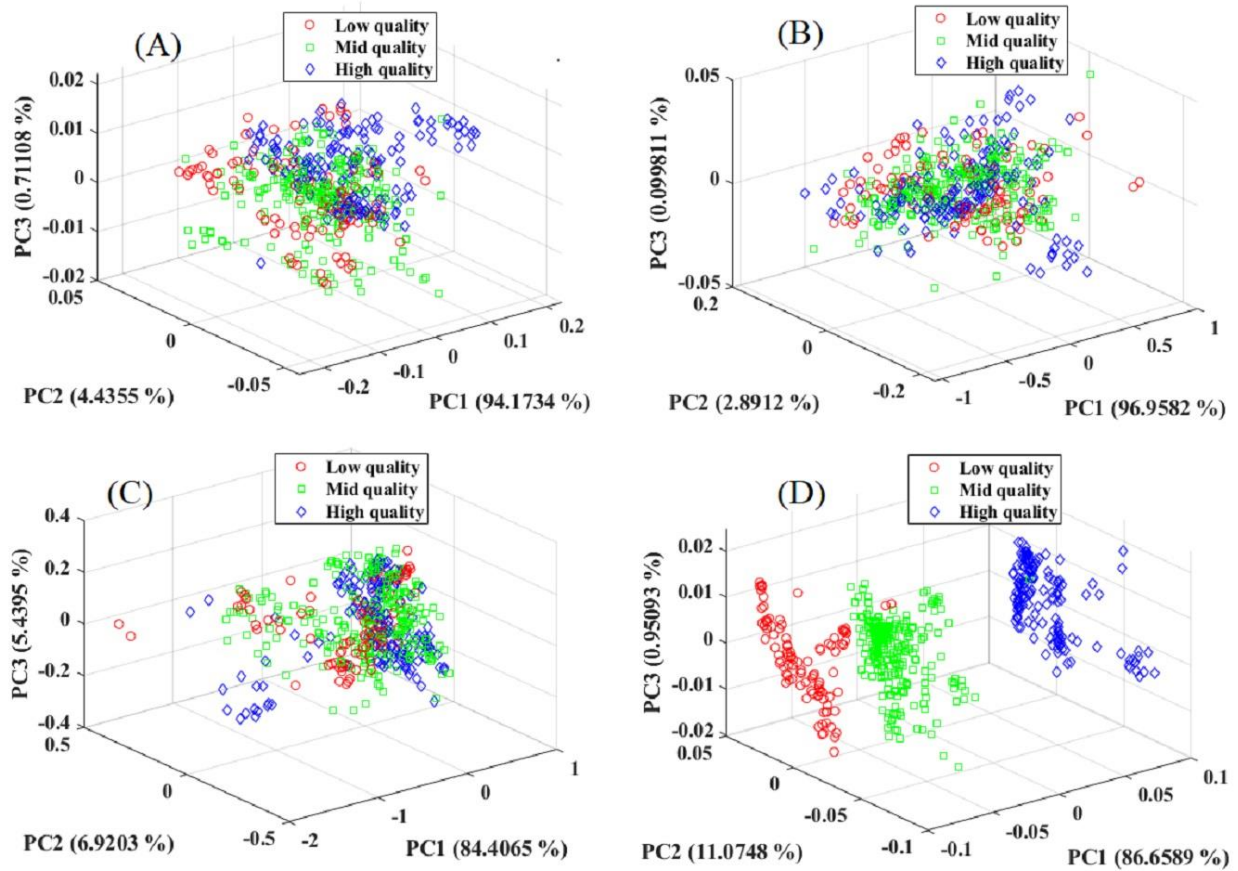
474

475

476

Figure 1. NIR spectra of rice samples; (A) = raw, (B) = mean of different quality categories, (C) = country of origin and (D) = imported versus

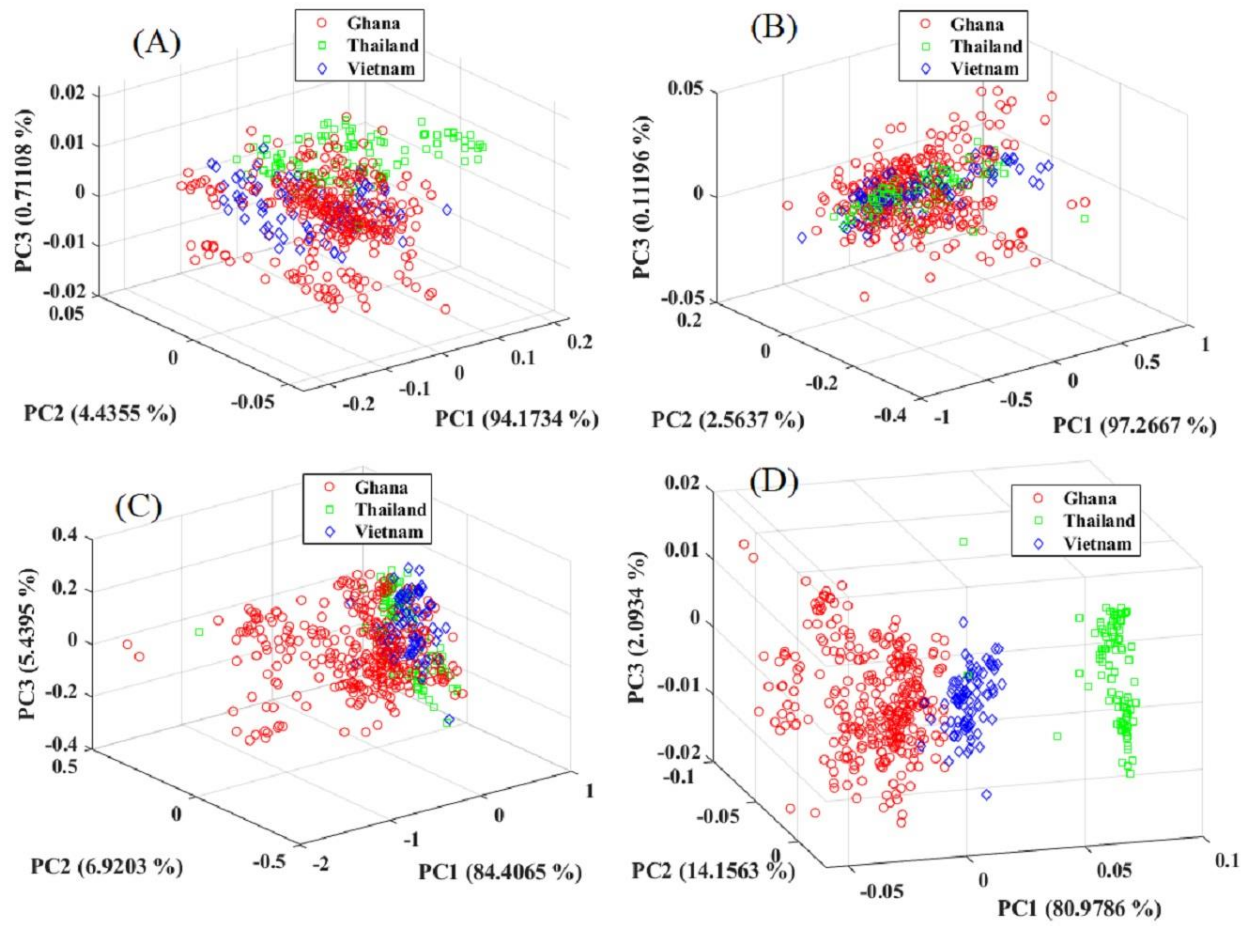
local



477

478

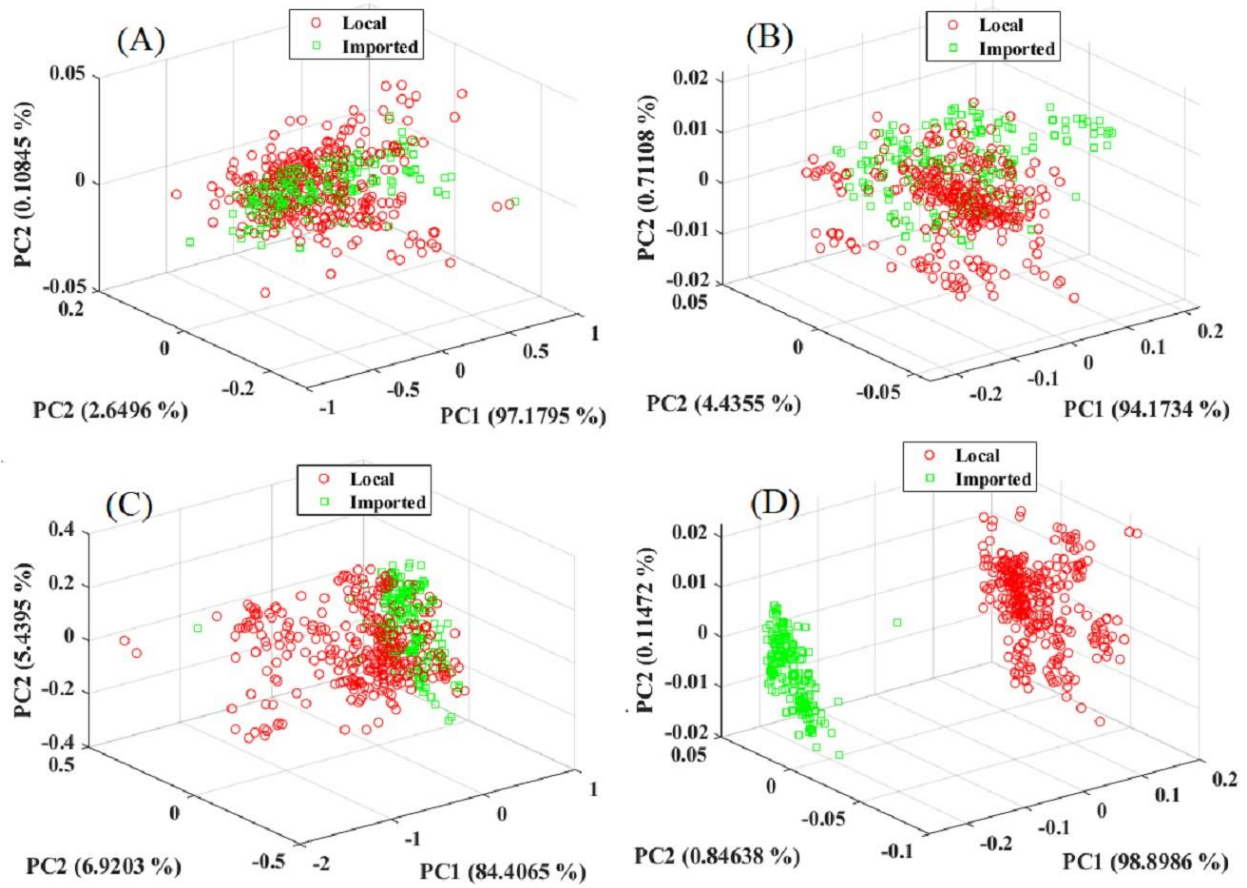
Figure 2. PCA score plot of rice sample of different quality preprocessed (A) MC, (B) DT, (C) SNV and (D) MSC.



479

480

Figure 3. PCA score plot of rice sample from different origin preprocessed (A) MC, (B) DT, (C) SNV and (D) MSC.



481

482

Figure 4. PCA score plot of rice sample local & imported preprocessed (A) MC, (B) DT, (C) SNV and (D) MSC.