

Classification trees based on infrared spectroscopic data to discriminate between genuine and counterfeit medicines

E. Deconinck¹, P.Y. Sacré^{1,2}, D. Coomans^{3,4}, J. De Beer^{1*}

¹Division of food, medicines and consumer safety, Section Medicinal Products, Scientific Institute of Public Health (IPH), J. Wytmansstraat 14, B-1050 Brussels, Belgium

²Department of Analytical Pharmaceutical Chemistry, Institute of Pharmacy, University of Liège, Liège, Belgium.

³Dept. of Biostatistics and Medical Informatics, Vrije Universiteit Brussel, Laarbeeklaan 103, B-1090 Brussels, Belgium

⁴Dept. of Analytical Chemistry and Pharmaceutical Technology, CePhaR, Vrije Universiteit Brussel-VUB, Laarbeeklaan 103, B-1090 Brussels, Belgium

Abstract:

Classification trees built with the Classification And Regression Tree algorithm were evaluated for modelling infrared spectroscopic data in order to discriminate between genuine and counterfeit drug samples and to classify counterfeit samples in different classes following the RIVM classification system.

Models were built for two data sets consisting of the Fourier Transform Infrared spectra, the Near Infrared spectra and the Raman spectra for genuine and counterfeit samples of respectively Viagra[®] and Cialis[®].

Easy interpretable models were obtained for both models. The models were validated for their descriptive and predictive properties. The predictive properties were evaluated using both cross validation as an external validation set. The obtained models for both data sets showed a 100% correct classification for the discrimination between genuine and counterfeit samples and 83.3% and 100% correct classification for the counterfeit samples for the Viagra[®] and the Cialis[®] data set respectively.

Keywords: counterfeit medicines, PDE-5 inhibitors, infrared spectroscopy, Raman spectroscopy, classification, CART

*corresponding author: Jacques.Debeer@wiv-isp.be

Tel. +32 2 642 51 70

Fax. +32 2 642 53 27

1. Introduction

Due to the extension of the internet, counterfeit drugs represent a growing threat for public health in the developing countries but also more and more in the industrial world [1,2]. The European Agency for Access to Safe Medicines (EAASM) claims that about 50% of the medicines sold through non identified/recognized websites are counterfeit and that 10% of the market in the developing countries and about 1% of the European market is covered by counterfeits [3]. In Europe and the United States one of the most popular group of medicines bought through the internet are the phosphodiesterase type 5 (PDE-5) inhibitors, i.e. sildenafil citrate (Viagra[®]), tadalafil (Cialis[®]) and vardenafil hydrochloride (Levitra[®]).

The World Health Organization (WHO) [4] defines a counterfeit drug as: “one which is deliberately and fraudulently mislabelled with respect to identity and/or source. Counterfeiting can apply to both branded and generic products and counterfeit products may include products with the correct ingredients or with the wrong ingredients, without the active ingredients, with insufficient active ingredient or with fake packaging.”

Even if this is the internationally accepted definition of a counterfeit medicine, it does not apply to the majority of the illegal products encountered on the European market, since they do not copy the packaging and brand names of the genuine products. Therefore it was chosen to follow the classification proposed by the Dutch National Institute for Public Health and the Environment (RIVM) [5]. This classification (table 1) distinguishes counterfeits, which appearance corresponds to the one of the genuine products, and imitations, which do not. Most of these imitations originate from Asia, where European and American patents are not recognized.

In literature several analytical techniques were proposed to discriminate between counterfeit and genuine medicines. Savaliya et al. proposed to use HPLC and LC-MS for the screening of aphrodisiacs on the Indian market [6]. Infrared-spectroscopy showed to be a valuable instrument for the identification of counterfeit medicines. Y. Roggo et al. [7] and Vajna et al. [8] made use of Raman spectroscopy for the identification of pharmaceutical tablets and investigate their structural differences, while de Veij et al. [9] used Raman spectroscopy to detect counterfeit Viagra[®]. Storme-Paris et al. [10], De Peinder et al. [11] Puchert et al. [12] and Lopes et al. [13,14] all demonstrated the usefulness of near infrared spectroscopy in the distinction of genuine and counterfeit medicines. Other techniques found for this purpose are colorimetry [15,16], TLC [17], NMR [18,19] and X-ray powder diffraction [20]. An overview of all these techniques can be found in reference [21].

Table 1: Classification of illegal medicines according to the RIVM [5]

Main category	Subcategory	Inclusion and exclusion criteria
Counterfeit	Professional	Appearance in conformity with genuine medicine; Content of correct API within 90 - 110 % of declared value; No other APIs; not genuine medicine.
	Non-professional	Appearance in conformity with genuine medicine; Content of correct API outside 90 - 110 % of declared value; No other APIs.
	Mixed	Appearance in conformity with genuine medicine; Contains correct API and another, known API
	Fraudulent	Appearance in conformity with genuine medicine; Contains a different, known API.
	Analogue	Appearance in conformity with genuine medicine, Contains other, unapproved API
	Placebo	Appearance in conformity with genuine medicine; Does not contain APIs.
Imitation	Professional	Appearance not in conformity with genuine medicine; Content of correct API within 90 - 110 % of declared value; No other APIs.
	Non-professional	Appearance not in conformity with genuine medicine; Content of declared API outside 90 - 110 % of declared value; No other APIs.
	Mixed	Appearance not in conformity with genuine medicine; Contains declared API and another API.
	Fraudulent	Appearance not in conformity with genuine medicine; Contains an undeclared API.
	Analogue	Appearance not in conformity with genuine medicine; Contains other, unapproved API
	Placebo	Appearance not in conformity with genuine medicine; Does not contain APIs.

One thing all these techniques have in common is that they generate a huge amount of data, which is often difficult to interpret in order to see differences between the different samples and to determine the cause of the differences. The majority of the authors make use of explorative chemometric tools to visualise the differences in the data obtained for the different samples: Principal Component Analysis, Partial Least Squares, Projection Pursuit, Multiple Factor Analysis and clustering techniques as hierarchical clustering, Generative topographic mapping and auto-associative multivariate regression trees are examples of methods that were and can be used for such purpose [9,22-24]. Even if some of the applied methods could be able to give a model with predictive ability, only a few authors created a model able to predict if a sample is counterfeit or not. Storme-Paris et al. [10] applied SIMCA to obtain a predictive model. The SIMCA model is a PCA-based model and the interpretation of the predictions is not always clear.

In this paper we evaluated the use of Classification And Regression Trees (CART) to build an easily interpretable predictive model to distinguish between counterfeit and genuine

medicines and to classify the counterfeit samples based on the RIVM definitions [5]. The first aim was to discriminate between genuine and counterfeit samples with at least 0% of genuines classified as counterfeits. The secondary aim was to be able to classify the counterfeit samples in their respective RIVM classes or at least get an idea about the type of the class.

To do so the data acquired by Sacré et al. [24] was used. The different samples were classified, following their visual aspects and the results obtained after identification and dosage of the active components, applying the RIVM classification system. The classes were used as response variable, while the spectroscopic data (FT-IR, NIR and Raman) were used as descriptive variables.

The results were compared to the results obtained by Sacré et al. [24]. In this paper the exploratory chemometric tools PCA and PLS were used for discrimination purposes. The results obtained with CART were also compared with the ones obtained with a more classic discrimination method, k-nearest neighbours (*k*NN). Due to the limited number of samples of some RIVM classes, SIMCA was not applied, since this will result in non representative classification rates for the counterfeit samples. This is due to the fact that in SIMCA each class is modelled separately.

2. Theory

2.1. Classification And Regression Trees (CART)

CART is a non-parametric statistical technique, developed by Breiman et al. [25] in 1984, which is able to solve classification (categorical dependent variables) as well as regression problems (continuous dependent variables). In both cases the method builds a decision tree, describing a response variable as a function of different explanatory variables (figure 1).

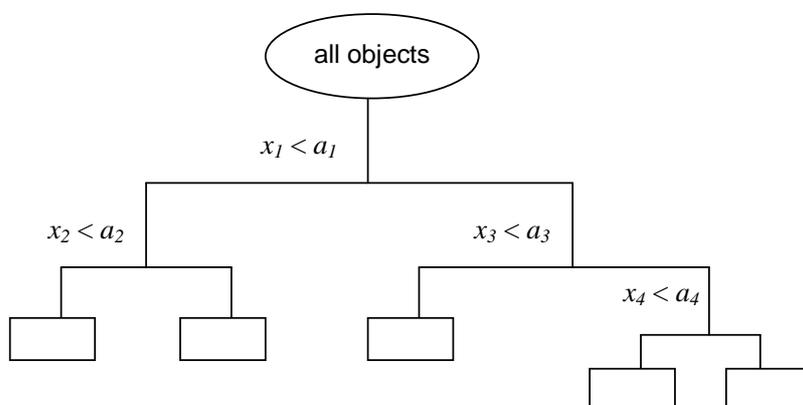


Figure 1: General structure of a CART-model. x_i = selected split variable, a_i = selected split value

A CART analysis generally consists of three steps. In a first step the maximum tree is built, using a binary split-procedure. The maximum tree is overgrown and closely describes the training set, usually resulting in overfitting. In a second step this overfitted model is pruned. This procedure results in a series of less complex trees, derived from the maximum tree. In the third and final step the optimal tree is selected using a cross validation procedure [25-27].

2.1.1. Building the maximum tree

The maximum tree is built using a binary split procedure, starting at the tree root, consisting of all objects in the training set. In every step of the procedure a mother group is considered and split into two daughter groups. The split is chosen in such a way that the impurity of the daughter groups is lower than that of the mother group. This means that the daughter groups become more homogeneous in the response variable (class numbers). In the following step each daughter group is considered a mother group. Every split is defined by one value of one explanatory variable. For continuous explanatory variables the splits are defined by " $x_i < a_j$ " where x_i is the selected variable and a_j its split value [25-27].

To choose the most appropriate variable and split value, CART uses an algorithm in which all descriptors and all possible split values are considered. The split resulting in the highest decrease in impurity between the mother group (t_p) and the daughter groups (t_L and t_R) is selected. Mathematically this is expressed as:

$$\Delta i(S, t_p) = i_p(t_p) - p_L i(t_L) - p_R i(t_R) \quad (\text{eq 1})$$

where i is the impurity, S the candidate split value, and p_L and p_R the fractions of the objects in the left and right daughter groups, respectively [25-27].

For classification trees the impurity can be defined by different split criteria [25]. The three commonly used split criteria are the Gini index, the Twoing index and the Information index. In this work CART models were built using the Gini index. The Information index and the Twoing index were not used since these measures were not found useful in solving the considered classification problem.

The Gini index is defined as

$$\Delta i = 1 - \sum_{j=1}^k (p_j(t))^2 \quad (\text{eq 2})$$

where $j = 1, 2, 3, \dots, k$ is the number of classes of the categorical response variable and $p_j(t)$ the probability of correct classification for class j at node t .

2.1.2. Tree pruning

The obtained maximum tree usually overfits the training set, therefore the model is pruned by successively cutting terminal branches. This procedure results in a series of smaller sub trees derived from the maximum tree. The different sub trees with the same complexity are then compared to find the optimal. This comparison is based on a cost-complexity measure $R_\alpha(T)$, in which both tree accuracy and complexity are considered [25-27]. For each sub tree complexity T it is defined as:

$$R_\alpha(T) = R(T) + \alpha |\tilde{T}| \quad (\text{eq 3})$$

with $R(T)$ the average within-node sum of squares, $|\tilde{T}|$ the tree complexity, defined as the total number of nodes of the sub tree, and α the complexity parameter, which is a penalty for each additional terminal node.

During the pruning procedure, α is gradually increased from 0 to 1 and for each value of α , the tree is selected which minimizes $R_\alpha(T)$. For a value of α equal to zero, $R_\alpha(T)$ is minimized by the maximum tree. By gradually increasing α a series of trees with decreasing complexity is obtained [25-27].

2.1.3. Selection of the optimal tree

From the obtained sequence of sub trees, the optimal has to be selected. The selection is usually based on the evaluation of the predictive error of the models using a cross validation procedure. In this paper a 10-fold cross validation procedure [26, 28] was used. The predictive error is then given as the overall misclassification rate for each of the sub trees [25]. The optimal model is the simplest model with a predictive error within one standard error (SE) of the minimal predictive error. This rule, generally referred to as the one SE-rule, allows the selection of a less complex model than the one with the minimal misclassification rate, without a significant loss of information and accuracy [25]. The algorithm allows the selection of all tree complexities. If previous knowledge about the data set justifies it, one can deviate from the one SE-rule and select another tree as optimal model.

2.2. *k* Nearest Neighbours (*k*NN)

The *k*-NN algorithm [28] was applied on the training set. The algorithm computes the minimal Euclidian distances between an unknown object and each of the objects of the training set.

For a training set of n samples, n distances are calculated. Then it selects the k nearest objects to the unknown one. The unknown object is classified in the group to which the majority of the k objects belong. The number of nearest neighbours is optimised using a cross validation procedure. The main advantages of this method are its mathematical simplicity and the fact that it is free from statistical assumptions.

3. Methods and materials

3.1. *Data*

The data for the Viagra[®] like samples consists of the Fourier-transformed infrared, the near-infrared and the Raman spectra for 55 counterfeit samples and 9 genuine samples. For the Cialis[®] like samples the data consists of the same type of spectra for 39 counterfeit and 4 genuine samples. All spectra were measured in triplicate. For more details about how this data was acquired we refer to Sacré et al. [24]. During the study of Sacré et al. it was seen that the variability in spectral data between the genuine samples is very low. The limited number of genuine samples should therefore be enough to represent the genuine class in the models. The more because CART defines the classes based on the improvement of homogeneity from mother to daughter leaves and therefore should isolate the small class of genuine relatively early in the building of the tree.

All counterfeit and imitation samples were donated by the Federal Agency for Medicines and Health Products in Belgium (AFMPS/FAGG). One batch of each dosage of genuine Viagra[®] (25mg, 50mg and 100mg) was kindly provided by Pfizer SA/NV (Belgium). Eli Lilly SA/NV (Benelux) kindly provided one batch of each commercial packaging (10 mg and 20 mg) of genuine Cialis[®]. Two other batches of each dosage of the genuine products were purchased from local pharmacies in Belgium.

The counterfeit samples were classified following the classification proposed by the RIVM [5]. The classification for both data sets is given in table 2

Table 2: Composition of the data sets in function of the RIVM classes [5]

Main category	subcategory	category number	number of Viagra [®] like samples	number of Cialis [®] like samples
Counterfeit	Professional	1	1	0
	Non-professional	2	3	0
	Mixed	3	1	5
	Fraudulent	4	0	0
	Analogue	5	0	0
	Placebo	6	0	0
Imitation	Professional	7	45	27
	Non-professional	8	4	2
	Mixed	9	1	5
	Fraudulent	10	0	0
	Analogue	11	0	0
	Placebo	12	0	0
Genuine		0	9	4

3.2. Data preprocessing

The data preprocessing was performed using the HoloREACT[™] software (Kaiser Optical Systems, USA, version 2.3.5). For NIR and FT-IR spectroscopy, the three spectra of a sample were normalized and averaged. For Raman spectroscopy, the three spectra of a sample were baseline corrected using the Pearson's method [29].

In order to evaluate the predictive ability of the models the Viagra[®] and Cialis[®] data sets were divided in training and test sets. It was chosen that the test sets would contain about 20% of the samples. The division in test- and training set was performed using the Duplexx algorithm [30]. This algorithm starts by selecting the two samples with the highest Euclidean distance in the data space for a first set. The next two samples with the highest Euclidean distance are selected for a second set. The procedure continues by selecting iteratively pairs of samples for the first and the second set. The second set was chosen as test set, while the other set combined with the samples not selected by the algorithm formed the training set. During selection of the test sets it was made sure that one genuine sample was selected for the test set, while the others were kept in the training set. This was necessary due to the limited number of genuine samples in the data sets.

From the PCA plots [31] shown in figure 2 and 3 it can be seen that the test sets selected by the Duplexx algorithm cover quite well the data space of the data sets.

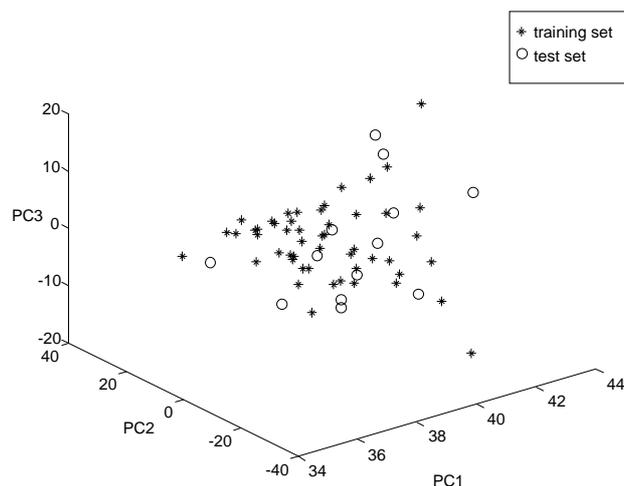


Figure 2: PCA plots representing the spread of the test set for the Viagra[®] data set over the data space.

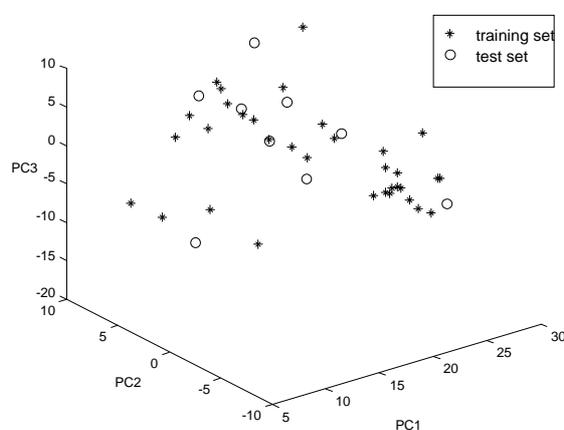


Figure 3: PCA plots representing the spread of the test set for the Cialis[®] data set over the data space.

3.3. Chemometrics

The data preprocessing and the modelling was performed using Matlab R2009b (The Mathworks, Natick, USA). The programming of the CART algorithm was done according to the original CART algorithm proposed by Breiman [25].

4. Results and Discussion

4.1. Viagra[®] like samples

For all data sets used the data was first autoscaled. The signals at the different wavelengths in the respective spectra were used as explanatory variables while the class numbers of the different samples were used as response. The class numbers were assigned to the different samples based on the classification proposed by RIVM (table 2).

The maximal tree was built and pruned. In a next step a 10-fold cross validation was carried out resulting in a graph representing the percentage misclassification as a function of the tree complexity.

4.1.1. Classification tree based on the FT-IR data

Figure 4 shows the graph of the percentage misclassification as a function of the tree complexity obtained with the FT-IR data. As can be seen from the figure the tree with complexity 2 has the smallest cross validation error. Since we have more than two classes and following the general rule that the optimal tree can have each complexity with a cross validation error within one standard deviation of the tree with the smallest error, the tree with complexity five was selected as optimal tree (figure 5). The cross validation error was 0.26 or 11.6%. Even if the cross validation error is quite high, it could be observed that during cross validation all genuine samples were classified correctly and none of the counterfeit or imitation samples were classified as genuine. This was confirmed during external validation, which means that the CART model is able to distinguish between genuine and counterfeit drugs, based on the FT-IR data. When focussing on the classification of the counterfeit or imitation samples over the different classes it was observed that 10 of the 12 counterfeit samples of the external test set were classified correctly while the 2 other were classified as non-professional imitations (class 8) in stead of professional imitations (class 7). Since the data set of the Viagra[®] like samples contains representatives of 7 classes, one should select a tree with minimal 7 leafs in order to be able to predict each class. Since our primary goal was to distinguish between counterfeit and genuine and the fact that we chose the optimal tree size based on the cross validation results, the leafs of the selected tree are not homogeneous for some counterfeit classes (class 1,3,9) and so the tree model is only able to give an indication about the sample type for these samples. The inhomogeneity for these classes in the model is probably due to the low number of samples belonging to these classes.

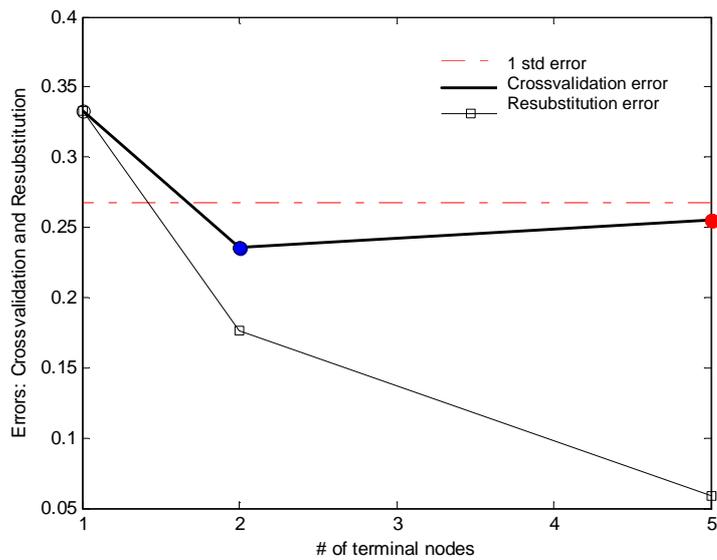


Figure 4: the cross validation error and the resubstitution error in function of the tree complexity for the FT-IR data

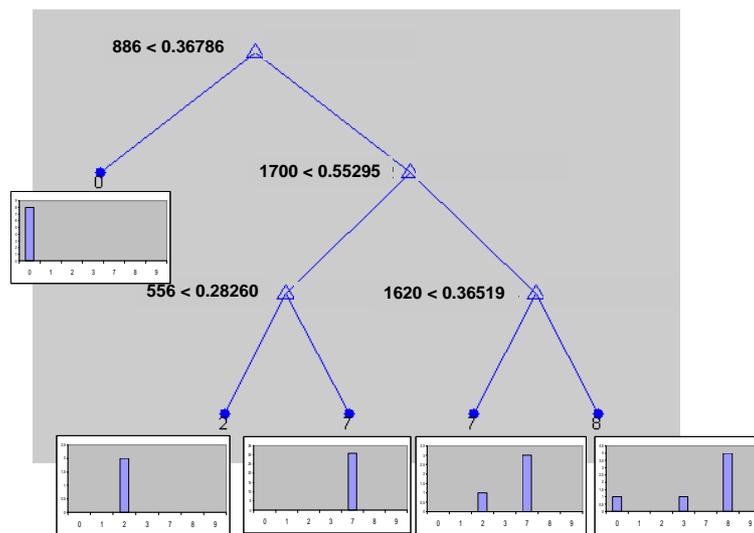


Figure 5: Classification tree based on the FT-IR data for the Viagra[®] data set. Each split is defined by the selected wavelength and its splitvalue. Each leaf is defined by the class number of the class most represented in the leaf and the graph gives the distribution (homogeneity) of the different samples in the leaf.

One of the main disadvantages of CART, used with spectral data is that it selects only one variable/wavelength to define each split. Since spectral data usually has a signal over a range of wavelengths this causes the difficulty to interpret the significance of the selected wavelengths. In fact for the CART model based on the FT-IR data the wavelengths of 886 cm^{-1} and 1700 cm^{-1} could be related to the concentration of sildenafil present in the samples. For the other selected wavelengths no logical explanation could be found.

4.1.2. Classification tree based on the NIR data

From a similar graph as shown in figure 4 it could be concluded that for the NIR data, the tree with complexity 4 has the lowest cross validation error and is the optimal tree (figure 6). The cross validation error was 0.37 or 14.4 %. Investigation of the tree shows that the leaf of the genuine samples is homogeneous and that during cross validation all genuine samples were classified correctly and that none of the counterfeit samples was classified as genuine. These results were also reflected during the external validation. During external validation, ten of the twelve counterfeit samples were classified correctly, while two professional imitations (class 7) were classified as non-professional counterfeits. From these results it can be concluded that the CART model based on NIR-data is comparable to the model obtained with the FT-IR data for its predictive and descriptive properties.

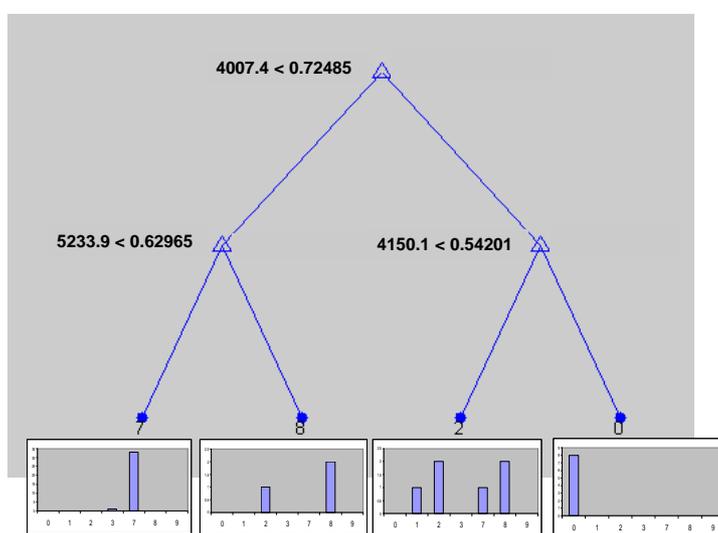


Figure 6: classification tree based on the NIR data for the Viagra[®] data set. Each split is defined by the selected wavelength and its splitvalue. Each leaf is defined by the class number of the class most represented in the leaf and the graph gives the distribution (homogeneity) of the different samples in the leaf.

Investigation of the selected variables revealed that the selected wavenumbers 4150.1 cm^{-1} and 5233.9 cm^{-1} correspond to specific peaks of the NIR spectrum of microcrystalline cellulose. It can therefore be stated that the discrimination of genuine and counterfeit samples by the CART model, based on NIR data, is partly based on the presence of different amounts of microcrystalline cellulose in the counterfeit and the genuine samples.

4.1.3. Classification tree based on the Raman spectroscopy data

For the Raman spectroscopic data the tree with complexity 3 was selected as the optimal tree (figure 7). The cross validation error was 0.25 or 11.5 %. Investigation of the tree and

the cross validation results show that the leaf representing the genuine samples (0) is homogeneous and that no genuine samples are classified in another leaf. Also during cross validation all genuine samples are correctly classified and none counterfeit sample is classified as genuine. The external validation reflects the same, all genuine are classified as genuine and no counterfeits are classified as genuine. When focussing on the classification of the counterfeit samples of the external test set, the model gives a correct classification for 8 of the 12 samples, which is significantly worse compared to the previous 2 models.

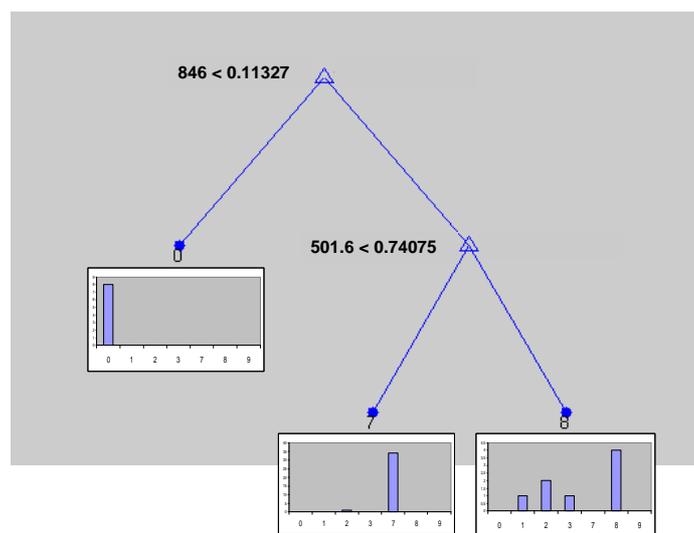


Figure 7: classification tree based on the Raman data for the Viagra[®] data set. Each split is defined by the selected wavelength and its splitvalue. Each leaf is defined by the class number of the class most represented in the leaf and the graph gives the distribution (homogeneity) of the different samples in the leaf.

The selected wavelengths could be linked to the presence of excipients. The wavelength of 501.6 cm^{-1} is a characteristic signal for the Raman spectrum of calcium hydrogenophosphate while 846 cm^{-1} is a wavelength corresponding to a signal characteristic for hydroxypropylcellulose. Calcium hydrogenophosphate is present in the genuine samples, but not or in different amounts in the counterfeit samples. Hydroxypropylcellulose on the other hand is not present in the genuine samples. Again it can be stated that the discrimination of the CART model is based on the presence of secondary components and excipients.

4.1.4. Classification trees based on the combination of the spectral data

Tree models were built using combinations of the data. One model was built using the FT-IR and the NIR data, one using the FT-IR and the Raman data, one using NIR and Raman and one using the combination of the three.

None of these models led to satisfying results, therefore it was decided not to discuss them in detail.

The models obtained with the combination of the FT-IR and the NIR data as well as the one based on the three types of spectral data, did not result in a satisfying model since a number of counterfeit samples was classified as genuine during both cross validation as external validation. Since the models based on each technique separately were able to make this distinction, it seems that combining the two data sets introduces noise in the model.

The other two models (combination of FT-IR and Raman and NIR and Raman respectively) were able to distinguish counterfeit from genuine samples, since no genuine sample were classified as counterfeit and no counterfeit as genuine, during both cross validation as external validation. During external validation it was seen that the models had higher misclassification rates (4/12 and 5/12) than the models discussed in section 4.1.1 and 4.1.2. From these results it has to be concluded that the combination of different types of spectral data, does not result in better models for the Viagra[®] data set.

4.2. *Cialis[®] like samples*

Exactly the same approach as described for the Viagra[®] data set was followed. The assignment of the class numbers, based on the RIVM classification, for the Cialis[®] data set is given in table 2.

4.2.1. *Classification trees based on the FT-IR and the NIR data*

From the graphs of percentage misclassification in function of the tree complexity the trees of complexity three and five were selected as optimal trees for respectively the FT-IR and the NIR data. Since cross validation and external validation showed that both models were not able to distinguish between genuine and counterfeit samples, these models are not further discussed.

4.2.2. *Classification tree based on the Raman spectroscopy data*

The tree of complexity five was selected as the optimal model obtained using the Raman spectroscopy data for the Cialis[®] data set (figure 8). A cross validation error of 0.62 or 28 % was obtained, which is high if it is compared to the errors obtained for the models for the Viagra[®] data set. Investigation of the leafs shows that the group of the genuine samples is homogeneous and that no genuine is classified with counterfeit samples. Also during cross validation all genuine samples are correctly classified and no counterfeit samples are classified as genuine. During external validation on the other hand it could be observed that

despite the fact that all genuine samples are correctly classified, two counterfeit samples are wrongly classified from which one is classified as genuine.

Three of the selected wavelengths (1109.1 cm^{-1} , 1119.0 cm^{-1} and 1120.2 cm^{-1}) could be linked to the Raman spectrum of lactose, an excipient present in both genuine as counterfeit samples. Probably the differences in amounts partly account for the discrimination.

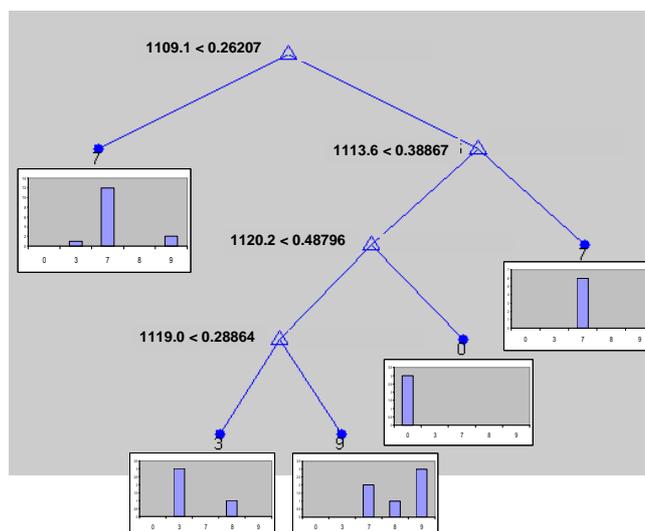


Figure 8: classification tree based on the Raman data for the Cialis[®] data set. Each split is defined by the selected wavelength and its splitvalue. Each leaf is defined by the class number of the class most represented in the leaf and the graph gives the distribution (homogeneity) of the different samples in the leaf.

4.2.3. Classification trees based on the combination of the spectral data

Tree models were built using a combination of the different types of spectral data. The only model able to distinguish between counterfeit and genuine samples was the one combining the NIR and the Raman spectroscopic data. The tree with complexity six was selected as the optimal tree (Figure 9). A cross validation error of 0.50 or 22.5 % was obtained. Investigation of the leafs showed that the leaf representing the genuine samples is homogeneous and that no genuine samples are classified with counterfeit ones. Also during cross validation and external validation all genuine are classified correctly and no counterfeit samples are classified as genuine. It has also to be mentioned that during external validation all counterfeit samples were classified correctly.

Three of the selected NIR wavelengths (4215.6 cm^{-1} , 5997.5 cm^{-1} and 6730.3 cm^{-1}) could be linked to excipients, present in the genuine Cialis[®] tablets. 4215.6 cm^{-1} corresponds to a characteristic peak of lactose, 5997.5 cm^{-1} to a characteristic peak of carmellose and 6730.3 cm^{-1} to one of microcrystalline cellulose. The two other wavelengths could not be linked to compounds, present or not in counterfeit samples.

Since three of the five wavelengths could be linked to excipients present in the genuine tablets, the discrimination is again probably due to the differences in amounts.

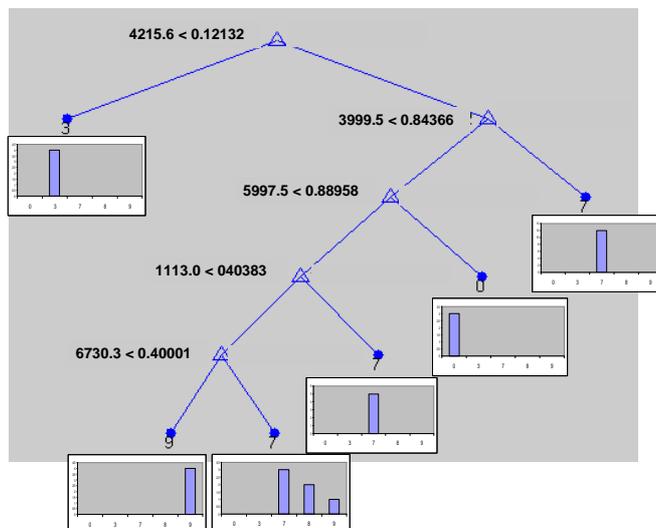


Figure 9: classification tree based on the combination of the NIR and the Raman data for the Cialis[®] data set. Each split is defined by the selected wavelength and its splitvalue. Each leaf is defined by the class number of the class most represented in the leaf and the graph gives the distribution (homogeneity) of the different samples in the leaf.

4.3. Results obtained with *k*NN

*k*NN was applied to the same data sets as described in sections 4.1 and 4.2. In a first step the optimal number of nearest neighbours was chosen, for each data set, using a leave-one-out cross validation procedure. The models were evaluated based on the correct classification rate (CCR) obtained during cross validation and the prediction of the external test set.

Two kind of *k*NN models were built, one using only the classes 0 for genuine samples and 1 for counterfeit samples and one using the different classes as used for the CART models. The reason is the fact that the theory of the algorithms (CART and *k*NN) are completely different. While CART will isolate the genuine group in a early phase of the model building and continue further classification of the counterfeits without influence of the genuine group, *k*NN will model the data set as a whole, resulting in a higher influence of the unbalanced numbers of samples in the different classes.

In general it can be stated that the results obtained with *k*NN are unsatisfying.

For the Viagra[®] dataset CCR values of 0.90 (optimal number of $k = 3$), 1.00 (optimal number of $k = 3$) and 0.9804 (optimal number of $k = 3$) were obtained for respectively the FT-IR, the NIR- and the Raman data, using cross validation when using the binary class (genuine vs. counterfeit) approach. Only the model based on the NIR data was able to discriminate between genuine and counterfeit, which was reflected by the 100 % CCR for both cross validation as external validation. In the models based on the FT-IR data and the Raman data at least one genuine was classified as counterfeit, which is unacceptable since it is a false positive. When using the different RIVM classes CCR values of 0.55 (optimal number of $k = 5$), 0.61 (optimal number of $k = 5$) and 0.43 (optimal number of $k = 9$) were obtained for respectively the FT-IR, the NIR- and the Raman data, using cross validation. For the external validation CCR values were obtained of 0.31, 0.23 and 0.23 respectively. The more in all models some originals were classified as counterfeit and some counterfeits as original. For the Cialis[®] dataset it was not possible to discriminate between original and counterfeit samples using k NN. Probably due to the fact that the number of genuine samples in this data set is too low for the algorithm.

5. Conclusions

The use of classification trees as easy interpretable models for the distinction of counterfeit and genuine drugs as well as the classification of counterfeit drugs, following the RIVM classification, was evaluated. All models were based on spectroscopic data. Models were built for two data sets, one consisting of the spectroscopic data for genuine and counterfeit samples of Viagra[®] and one consisting of the spectroscopic data for genuine and counterfeit samples of Cialis[®]. An overview of the obtained correct classification rates for the different models proposed is given in table 3. Only models with a 100% correct classification rate for the discrimination between genuine and counterfeit, during cross validation, are present in the table, since only these models are of interest.

For the Viagra[®] data set, two comparable models could be proposed, one based on the FT-IR data and one on the NIR data. The models have cross validation errors of 11.6% and 14.4% respectively and equal misclassification rates of 2/12 after external validation. Since both models are comparable the choice is based on the equipment present in the laboratories. For the both models the classification/discrimination could partly be explained by the differences in amounts of active substance (FT-IR model) and excipients (NIR model). The combination of the different types of spectroscopic data did not result in better models compared to the ones obtained with only FT-IR or NIR data.

Table 3: Overview of the prediction errors of the proposed models

	Data used	CCR genuine/counterfeit (cross-validation) (%)	Overall CCR	CCR genuine/counterfeit	Overall CCR
			(cross validation) (%)		
Viagra® like samples	FT-IR	100	88,4	13/13	11/13
	NIR	100	85,6	13/13	11/13
	Raman	100	88,5	13/13	9/13
Cialis® like samples	Raman	100	72	8/9	6/9
	Raman-NIR	100	77,5	9/9	9/9

For the Cialis® data set the best model was obtained by combining the NIR and the Raman spectroscopic data. All other models were not able to make the distinction between counterfeit and genuine samples. Only the model based on the Raman spectroscopy data was able to classify all genuine samples correctly, but also classified a counterfeit sample as genuine.

The proposed model has a cross validation error of 22.5%, but showed a 100% correct classification during external validation. Again the majority of the selected wavelengths could be linked to differences in amounts of excipients between the different samples.

The obtained results show that the application of CART to spectroscopic data result in easy interpretable models, which are able to discriminate between counterfeit and genuine drug samples and which are able to classify the counterfeit samples in their corresponding RIVM class and this with low misclassification rates, evaluated with an external test set. The classification of the counterfeits in their respective RIVM class, can allow an easy and fast evaluation of the risk for public health of a considered counterfeit sample.

Further it was shown that the results obtained with this tree based method are far better than the ones obtained with the more traditional discriminating method *k*NN. If the results of the CART models are compared to the results obtained with PLS by Sacré et al. [24], it can be seen that both methods are able to discriminate between genuine and counterfeit products, but that CART also allow a clear discrimination of the counterfeit samples in different classes. The more CART has the advantage to be able to classify the samples in a discriminating and easy interpretable model. Though the models are limited by the nature of the data set and they should be adapted and updated each time new samples/classes are encountered.

References

- [1] Weiss A., Buying prescription drugs on the internet: promises and pitfalls, *Cleveland Clinic Journal of Medicine* (2006)73, 282-288
- [2] Veronin M., Youan B.-B., Magic bullet gone astray: medications and the internet, *Science* (2004) 305, 481
- [3] European Alliance For Acces to Safe Medicines: www.eaasm.eu
- [4] WHO, sixty-second world health assembly item 12.9, counterfeit medical products, April 2009. http://aps.who.int/gb/ebwha/pdf_files/A62/A62_13-en.pdf
- [5] Venhuis B.J., Barends D.M., Zwaagstra M.E., de Kaste D., Recent developments in counterfeit and imitations of Viagra, Cialis and Levitra, RIVM Report 370030001/2007, Bilthoven, 2007
- [6] Savalyia A.A., Shah R.P., Prasad B., Singh S., Screening of Indian aphrodisiac ayurvedic/herbal healthcare products for adulteration with sildenafil, tadalafil and/or vardenafil using LC/PDA and extracted ion LC–MS/TOF, *Journal of Pharmaceutical and Biomedical Analysis* (2010) 52, 406-409
- [7] Roggo Y., Degardin K., Margot P., Identification of pharmaceutical tablets by Raman spectroscopy and chemometrics, *Talanta* (2010) 81, 988-995
- [8] Vajna B., Farkas I., Szabó A., Zsigmond Z., Marosi G., Raman microscopic evaluation of technology dependent structural differences in tablets containing imipramine model drug, *Journal of Pharmaceutical and Biomedical Analysis* (2010) 51, 30-38
- [9] de Veij M., Deneckere A., Vandenaabeele P., de Kaste D., Moens L., Detection of counterfeit Viagra with Raman spectroscopy, *Journal of Pharmaceutical and Biomedical Analysis* (2008) 46, 303-309
- [10] Storme-Paris I., Rebiere H, Matoga M., Civade C., Bonnet P.-A., Tissier M.H., Chaminade P., Challenging near infrared spectroscopy discriminating ability for counterfeit pharmaceuticals detection, *Analytica Chimica Acta* (2010) 658, 163-174
- [11] de Peinder P., Vredenburg M.J., Visser T., de Kaste D., Detection of Lipitor counterfeits: a comparison of NIR and Raman spectroscopy in combination with chemometrics, *Journal of Pharmaceutical and Biomedical Analysis* (2008) 47, 688-694
- [12] Puchert T., Lochmann D., Menezes J.C., Reich G., Near-infrared chemical imaging (NIR-CI) for counterfeit drug identification--a four-stage concept with a novel approach of data processing (Linear Image Signature), *Journal of Pharmaceutical and Biomedical Analysis* (2010) 51, 138-145
- [13] Lopes M.B., Wolff J.C., Investigation into classification/sourcing of suspect counterfeit Heptodintrade mark tablets by near infrared chemical imaging, *Analytica Chimica Acta* (2009) 633, 149-155

- [14] Lopes M.B., Wolff J.C., Bioucas-Dias J.M., Figueiredo M.A., Determination of the composition of counterfeit Heptodin tablets by near infrared chemical imaging and classical least squares estimation, *Analytica Chimica Acta* (2009) 641, 46-51
- [15] Amin A.S., Moustafa M.E., El-Dosoky R., Colorimetric determination of sildenafil citrate (Viagra) through ion-associate complex formation, *Journal of AOAC International* (2009) 92, 125-130
- [16] Rodomonte A.L., Gaudio M.C., Antoniella E., Lucente D., Crusco V., Bartolomei M., Bertocchi P., Manna L., Valvo L., Muleri N., Counterfeit drugs detection by measurement of tablets and secondary packaging colour, *Journal of Pharmaceutical and Biomedical Analysis* (2010) 53, 215-220
- [17] Vredenburg M.J., Blok-Tip L., Hoogerbrugge R., Barends D.M., de Kaste D., Screening suspected counterfeit Viagra and imitations of Viagra with near-infrared spectroscopy, *Journal of Pharmaceutical and Biomedical Analysis* (2006) 40, 840-849
- [18] Wawer I., Pisklak M., Chilmonczyk Z., ¹H, ¹³C, ¹⁵N NMR analysis of sildenafil base and citrate (Viagra) in solution, solid state and pharmaceutical dosage forms, *Journal of Pharmaceutical and Biomedical Analysis* (2005) 38, 865-870
- [19] Trefi S., Routaboul C., Hamieh S., Gilard V., Malet-Martino M., Martino R., Analysis of illegally manufactured formulations of tadalafil (Cialis) by ¹H NMR, 2D DOSY ¹H NMR and Raman spectroscopy, *Journal of Pharmaceutical and Biomedical Analysis* (2008) 47, 103-113
- [20] Maurin J.K., Pluciński F., Mazurek A.P., Fijałek Z., The usefulness of simple X-ray powder diffraction analysis for counterfeit control--the Viagra example, *Journal of Pharmaceutical and Biomedical Analysis* (2007) 43, 1514-1518
- [21] Martino R., Malet-Martino M., Gilard V., Counterfeit drugs: Analytical techniques for their identification, *Analytical and Bioanalytical Chemistry* (2010) 398, 77-92.
- [22] Deconinck E., van Nederkassel A.M., Stanimirova I., Daszykowski M., Bensaid F., Lees M., Martin G.J., Desmurs J.R., Smeyers-Verbeke J., Vander Heyden Y., Isotopic ratios to detect infringements of patents or proprietary processes of pharmaceuticals: two case studies, *Journal of Pharmaceutical and Biomedical Analysis* (2008) 48, 27-41.
- [23] Dumarey M., van Nederkassel A.M., Stanimirova I., Daszykowski M., Bensaid F., Lees M., Martin G.J., Desmurs J.R., Smeyers-Verbeke J., Vander Heyden Y., Recognizing paracetamol formulations with the same synthesis pathway based on their trace-enriched chromatographic impurity profiles, *Analytica Chimica Acta*. (2009) 655, 43-51.
- [24] Sacré P-Y, Deconinck E., De Beer T., Courselle P., Vancauwenberghe R., Chiap P., Crommen J., De Beer J.O., Comparison and combination of spectroscopic techniques for the detection of counterfeit medicines, *Journal of Pharmaceutical and Biomedical Analysis* (2010) 53, 445-53

- [25] Breiman L., Friedman J.H., Olshen R.A., Stone C.J., Classification and regression trees; Wadsworth & Brooks, Monterey, 1984.
- [26] Deconinck E., Hancock T., Coomans D., Massart D.L., Vander Heyden Y., Classification of drugs in absorption classes using Classification And Regression Trees (CART)-methodology, Journal of Pharmaceutical and Biomedical Analysis (2005) 39, 91-103.
- [27] Deconinck E., Zhang M.H., Coomans D., Vander Heyden Y., Classification tree models for the prediction of blood-brain barrier passage of drugs, Journal of Chemical Information and Modeling (2006) 46, 1410-1419.
- [28] Vandeginste B.G.M., Massart D.L., Buydens L.M.C., De Jong S., Lewi P.J., Smeyers-Verbeke J.: Handbook of Chemometrics and Qualimetrics-Part B, Elsevier Science, Amsterdam, 1997
- [29] Pearson G.A., A general baseline-recognition and baseline flattening algorithms, Journal of Magnetic Resonance (1977) 27, 265-272.
- [30] Snee R.D.: Validation of regression models: Methods and examples. Technometrics (1977) 19, 415-428.
- [31] Massart D.L., Vandeginste B.G.M., Buydens L.M.C., De Jong S., Lewi P.J., Smeyers-Verbeke J.: Handbook of Chemometrics and Qualimetrics-Part A, Elsevier Science, Amsterdam, 1997