Prediction of pellet quality through machine learning techniques and near-infrared spectroscopy

Prediction of pellet quality through machine learning techniques and near-infrared spectroscopy

Manuela Mancini[b], Alex Mircoli[a,*], Domenico Potena[a], Claudia Diamantini[a], Daniele Duca[b], Giuseppe Toscano[b]

[a] *Department of Information Engineering, Università Politecnica delle Marche, Ancona, Italy*

[b] *Department of Agricultural, Food and Environmental Sciences, Università Politecnica delle Marche, Ancona, Italy*

Abstract

In recent years, pellet has received increasing attention among other biofuels due to its low storage costs and high combustion efficiency. The traceability of pellet quality along the entire supply chain is a critical issue, since fraudulent behaviours, such as the replacement with lower quality pellet, may both cause an economic damage and harm consumers' health. Traditionally, pellet quality is evaluated through laboratory analysis, which is costly and time-consuming. To overcome these limitations, in this work we define a methodology for quick and low-cost evaluation of pellet quality, which may be used along the entire supply chain. The proposed technique is based on the classification of pellet spectra through machine learning techniques. Spectra are obtained by means of a near-infrared (NIR) spectrophotometer, which is a relatively cheap instrument of small dimensions (even portable) that is suitable for on-site analysis at any phase of the supply chain. We propose two different approaches, namely an automatic classification of pellet, which does not require laboratory analysis, and a semi-automatic approach, that increases the overall accuracy but requires laboratory analysis for uncertainly classified samples. We validate the methodology by performing several experiments on real-world data, by training different machine learning algo-

*Corresponding author

Email addresses: m.mancini@univpm.it (Manuela Mancini), a.mircoli@univpm.it (Alex Mircoli), d.potena@univpm.it (Domenico Potena), c.diamantini@univpm.it (Claudia Diamantini), d.duca@univpm.it (Daniele Duca), g.toscano@univpm.it (Giuseppe Toscano)

rithms and evaluating the impact of several transformations introduced to reduce the scattering effect, which is a well-known issue related to NIR data.

Keywords: pellet quality, near-infrared spectroscopy, machine learning, supply chain, classification of ash content

1. Introduction

Nowadays, renewable energy is becoming increasingly important as a mean to meet the growing global energy demand[1], while keeping the impact on the environment under control. Among the different types of renewable energies, energy from biomass sources has proven to be of strategic importance. According to the International Energy Agency (IEA), it accounts for 9% of global primary energy supply (IEA, 2017). In particular, pellet turns out to be one of the most competitive biomass sources (Magelli et al., 2009; Mola-Yudego et al., 2014; Selkimaki et al., 2010). Although pellet can be made up of different biomass materials, traditionally the most used are the woody materials from wood processing industries, wood shavings and wood chips (Nielsen et al., 2009). Pellet is created by the pelletizing process, which consists in condensing raw materials under heat and pressure. The result is a product with cylindrical shape, high heating value and low moisture content. These features make pellet advantageous in comparison to other biomasses, as they allow to obtain a clean burning and a reduction of produced ashes. In addition, the high fuel density of pellet results in a reduction of transportation and storage costs (Mola-Yudego et al., 2014; Selkimaki et al., 2010). Pellet can be used both for residential heating and industrial use (co-generation plants). In both cases its quality plays an essential role. Pellets with high quality and more constant properties make easier to regulate the combustion process, avoiding technical problems, reducing maintenance costs and ensuring high efficiency (Filbakk et al., 2011; García et al., 2019). This is particularly true for small-scale boilers and stoves, since industrial boilers are equipped with more advanced flue gas cleaning, combustion and process control systems (García et al., 2019).

According to the technical standard EN ISO 17225-2, pellet is divided into three quality classes (A1, A2 and B) on the basis of qualitative features

[1] According to the International Energy Agency the global energy demand grew by 2.3% in the 2018 - https://www.iea.org/geco/

and chemical-physical parameters. The A1 and A2 classes guarantee quality requirements that allow pellet to be used in residential devices, while pellet belonging to B class can be used only for industrial purposes. Out of specification (OOS) pellets are of poor quality, but can be still used in special industrial boilers equipped with more advanced flue gas cleaning and process control systems able to handle high ash content. With regard to the chemical parameters, ash content is one of the most relevant to be monitored on biomass samples (Duca et al., 2014). In fact, ash content is related to other critical elements like sulphur, chlorine and potassium and gives useful indications about problems for combustion devices (Toscano et al., 2013, 2016; Monti et al., 2008). In particular, the technical standard assigns samples with ash content less than 0.7% to A1 class, with ash content between 0.7% and 1.2% to A2 class and with ash content between 1.2% and 2.0% to B class.

Given the increasing wood pellet imports in Europe over the years (Sun and Niquidet, 2017), it is fundamental to guarantee the traceability of its quality. In fact, since the pellet supply chain involves several actors from production to retail, there is high risk of fraudulent behaviours (e.g., substitution with lower quality pellet) that may affect the quality of the combustion and hence expose consumers to health risks. Currently, ash content can be determined by sending a pellet sample to a laboratory for specific analysis, which requires some days. Nevertheless, biomass has a very high variability and chemical complexity, which imply that the results of the analysis of a single sample may not be representative of the entire batch. Furthermore, in order to trace the quality of pellet throughout the entire supply chain, different analysis should be performed at each phase. Hence, determining pellet quality takes a long time, which results in additional costs in terms of pellet immobilization at each phase of the supply chain.

In this paper we propose a cheap and quick methodology to assign pellet samples to a quality class. To this end, we use machine learning techniques to classify pellet samples, represented as spectra obtained by means of near-infrared (NIR) spectrophotometer. NIR spectroscopy is a rapid and low-cost analysis technique, and hence can be used in every phase of the supply chain. Indeed, since the NIR spectrophotometer is small (even portable), it can be used directly on a pellet batch, avoiding the delay due to laboratory analysis (i.e., sample shipping, sample preparation and analysis). Furthermore, since acquisition time needs few seconds, several spectra from different samples can be easily gathered, reducing the issue of sample representativeness. In

details, we propose two different approaches: i) automatic classification of pellet on the basis of ash content, which does not require laboratory analysis, and ii) semi-automatic approach, which increases the overall accuracy but requires laboratory analysis for uncertainly classified samples. Furthermore, we evaluate several preprocessing techniques which aim at reducing two well-known issues related to NIR method, namely the scattering effect and the high dimensionality of data.

The rest of the paper is structured as follows: in Section 2 we present some related work on the evaluation of pellet characteristics through NIR spectroscopy. Section 3 introduces the methodology for preprocessing and classification of pellet, while Section 4 presents the results of experiments on real-world data. Finally, Section 5 draws conclusions and discusses future work.

2. Related Work

Different studies have investigated the applicability of NIR spectroscopy for the analysis of the quality of solid biofuels. NIR method has been examined for the prediction of both qualitative (i.e., origin and source) and quantitative parameters of biomass, as requested by the technical standard EN ISO 17225.

NIR method has been already investigated for predicting the ash content on different biofuels. In Fagan et al. (2011), authors tried to predict the main chemical-physical properties of two dedicated bioenergy crops using NIR spectroscopy. In particular, Partial Least Squares (PLS) regression models were developed for the prediction of moisture, ash, carbon and nitrogen content with poor results for both ash and carbon content. In Gillespie et al. (2015), a hyperspectral imaging instrument is used for the prediction of the same quantitative parameters directly on pellet samples. Even in this case the performance of the PLS model for the prediction of ash content resulted to be poor. In Maranan and Laborie (2008), NIR spectra of different hybrid poplar clones samples have been acquired and their chemical characteristics predicted using PLS models.

NIR spectroscopy has been used also for real-time monitoring of pellet quality. In Lestander et al. (2009), authors studied the possibility of using online NIR spectroscopy directly in the pelletizing process for predicting moisture content, sawdust blends and energy consumption of the pellet

press, with the aim of optimizing the pellet production process. In a successive study, the authors also examined the opportunity of getting information about species composition and moisture content of the dried wood particles (Lestander et al., 2012). The possibility of predicting moisture content, gross calorific value and ash content in real time by analyzing wood samples placed in a rotating cup through NIR spectroscopy was studied in Lestander and Christofer (2005). A reasonably good model for the prediction of ash content was developed using bi-orthogonal partial least squares regression (BPLS).

It is to be noted here that all the mentioned studies have used regression algorithms for the prediction of the ash content of pellet samples. Considering the poor results of the regression models for the ash content prediction and the standard EN ISO 17225-2 which defines three quality classes for pellet on the basis of ash content values, the use classification models could be a good alternative to predict pellet quality. To the best of our knowledge, the use of classification algorithms in literature is focused on the evaluation of pellet shape and density, which is performed through radial basis function networks (Kusumoputro et al., 2013) or adaptive neuro fuzzy inference systems (ANFIS) (Sutarya and Kusumoputro, 2011), and to the detection of origin and source of the material (Santoni et al., 2015; Sandak et al., 2011; Espinoza et al., 2012).

In order to guarantee the product's complete traceability, the analysis of pellet quality should be extended to the entire supply chain, which is usually long and has different analysis requirements among the various steps. Nevertheless, currently there is no work focused on such critical aspect. In fact, existing studies (Quddus et al., 2017) examined only the possibility of developing optimization models for the design and management of the biofuel supply chain.

3. Methodology

It is fundamental to get information about pellet quality not only at the end of the pelletizing process, when the product is ready to be sold, but at each step of the supply chain, in order to monitor and assess the quality of the entire process. To enable this control, we define a methodology that consists in analyzing pellet samples by means of spectroscopy techniques and classifying the resulting spectral data through machine learning techniques. The methodology includes different steps, as depicted in Figure 1. First, near-infrared data gathering is performed on a pellet sample; then, the re-

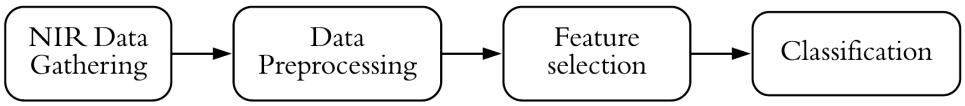

Figure 1: The proposed methodology for pellet classification

sulting spectra are treated with preprocessing techniques aimed at reducing scattering effects and data dimensionality. Finally, spectral data are classified through machine learning algorithms.

A detailed description of each methodology step is presented in the following subsections, along with the discussion of the main issues.

3.1. NIR Data Gathering

The main goal of this step is the acquisition of near-infrared spectra from different pellet samples.

In NIR spectroscopy, molecular chemical bonds are struck by near-infrared radiation, which causes vibrations at different energy level on the basis of the molecular structure and chemical composition of the material. The spectrophotometer returns a spectrum, where for each NIR wavelength the corresponding energy absorption value is reported (Pasquini, 2003).

Different NIR spectrophotometers can be used for spectral analysis. Since the on-line quality control of pellet can be performed directly in the production line, it is important for the instrument to be robust, so that its optical parts are not damaged by environmental humidity or dust. Therefore, the technology employed for the wavelength selection should be based on robust solutions, as fixed dispersive optics or sensor arrays, where no moving parts are present. In this work, spectral measurements have been performed in reflectance using an online NIR spectrophotometer (NIR-Online; Buchi Labortechnik AG, Flawil, Switzerland). The instrument is based on diode-array technology and is equipped with a rotating module (X-Rot module; Buchi Labortechnik AG, Flawil, Switzerland) placed below the spectrophotometer and simulating the production line in a pellet plant or the input material in a power plant. Each spectrum is an average of 32 successive scans; all the scans have been performed at room temperature (18-20 °C). The wavelength range is from 400 to 1,700 nm and the spectral resolution is 5 cm^{-1}, resulting in 261 absorption values.

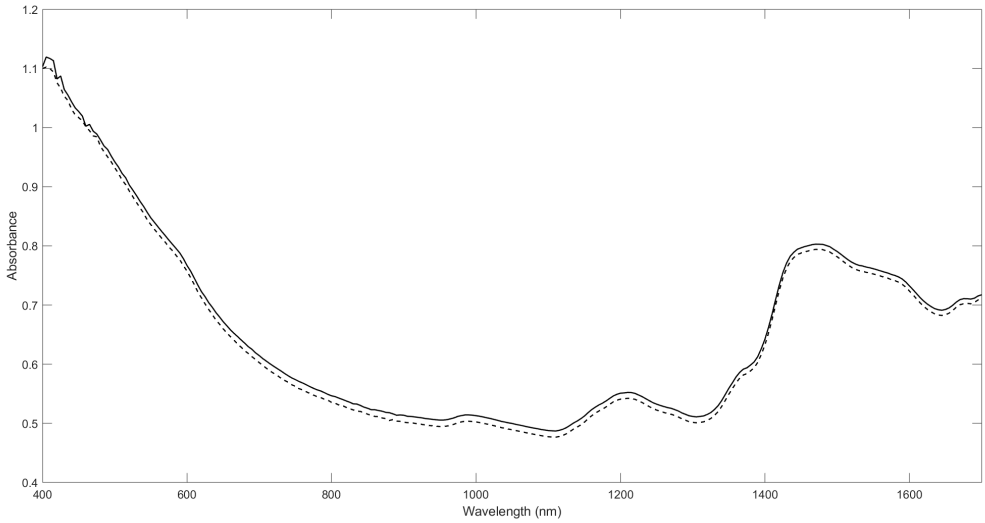

Figure 2: Two spectra (dotted and solid lines) obtained repeating the NIR analysis on the same randomly selected pellet sample.

Since the pellet usually has a cylindrical shape, data gathering through NIR spectrophotometer could be affected by scattering effect. This is a typical physical phenomenon that happens during spectra acquisition from solid samples, consisting in the deviation of light from a straight trajectory. Scattering effect leads to different spectra when the NIR analysis is repeated on the same sample, as shown in Figure 2. The average distance Δ between two different spectra S_1 and S_2 can be calculated through the formula $\Delta = \frac{1}{N}\sum_{i=1}^{N}\sqrt{(S_1(f_i) - S_2(f_i))^2}$, where $S_1(f_i)$ and $S_2(f_i)$ are the absorbance values at frequency f_i of S_1 and S_2 respectively. The value of Δ of the two spectra in Figure 2 is $1.03 \cdot 10^{-2}$.

In order to take into account the scattering effect, we apply two strategies: a physical transformation of the sample and a transformation of the spectrum. The first transformation consists in stabilizing the material at 40°C for 24 hours and grinding it below 1 mm of particle size. This transformation is compliant to UNI EN 14780:2011:"Methods for sample preparation", and guarantees that all samples reach similar moisture content and the spectral differences are related only to chemical differences. The transformation of

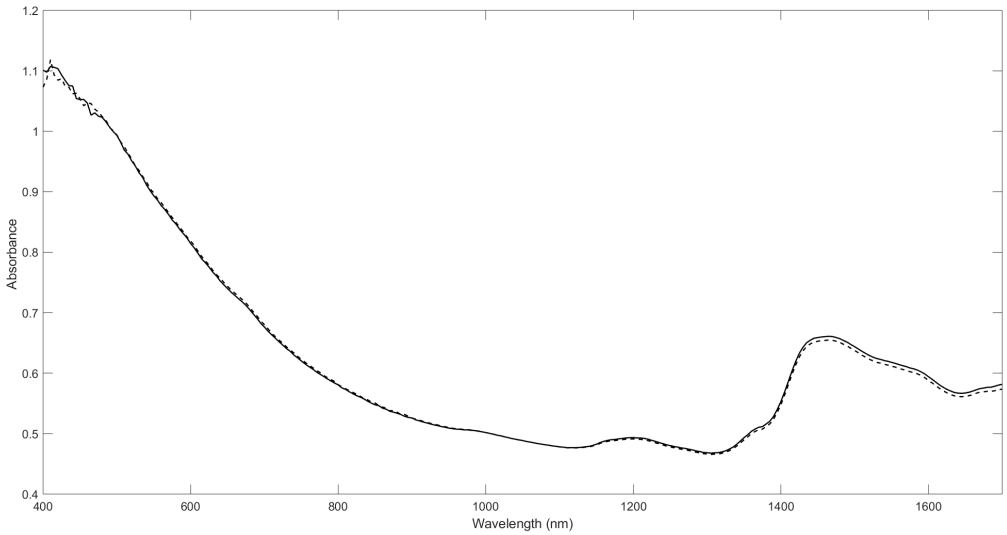

Figure 3: Two spectra (dotted and solid lines) obtained repeating the NIR analysis on the same randomly selected ground pellet sample.

the spectrum is based on the use of derivatives and will be described on Subsection 3.2.

Figure 3 shows two spectra obtained by repeating the NIR analysis after having ground the same sample used in Figure 2. In this case, the scattering effect is reduced; indeed, the average distance between the two spectra is $\Delta = 3.5 \cdot 10^{-3}$, which is lower than the one computed on the two spectra in Figure 2.

Hence, data acquisition through NIR spectrophotometer is performed on both original and, after the sample preparation, ground pellet samples. Hereafter we refer to the two sample sets as *pellet* and *ground pellet*.

Finally, in order to define the training set to be used in the classification step of the proposed methodology, we assign each sample to a quality class. To this end, ash content is calculated according to EN ISO 18122:2015. More in detail, the ash content can be determined using a thermo-gravimetric analyzer (mod. 701 Leco). The ground pellet samples are incinerated using a muffle furnace at the controlled temperature of 550 ± 10°C, then the residues mass is used to calculate the percentage of ash content air dried (Ac). Samples with ash content less than 1.2% are assigned to class A, samples with

ash content between 1.2% and 2.0% are assigned to class B and samples with ash content greater than 2.0% are assigned to the out-of-specification (OOS) class that can be used only for limited industrial applications with a reduced economic value. Unlike the technical standard EN ISO 17225-2, we combine the two classes A1 and A2 into the class A since they can be used in the same combustion devices.

3.2. Data Preprocessing

As introduced in 3.1, the transformation of spectrum is a way to reduce the scattering effect. In particular, the use of first and second derivatives is a widely adopted preprocessing technique in spectroscopy to reduce offset difference between spectra (Rinnan et al., 2009)(Rinnan, 2014).

The Savitzky-Golay filter (Savitzky and Golay, 1964) is a method used for smoothing and deriving spectra with the aim of decreasing the detrimental effect on the signal-to-noise ratio of the spectra. Basically, the method computes several convolutions by fitting a low-degree polynomial on successive sub-sets of adjacent frequencies. The fitted polynomial is used to calculate the first and second derivatives at the central point of each sub-set. The procedure is applied for all the points of the spectra subsequently.

As an example, Figure 4 shows the same spectra reported in Figure 2 pretreated with second derivative (Savitzky-Golay filter with 13 smoothing points and 2nd order polynomial). It is to be noted that the second derivative removes the additive and slope effects of the scattering, resulting in a reduction of the offset between the spectra. Indeed, the average distance between the two spectra in Figure 4 is $\Delta = 2 \cdot 10^{-5}$, which is much lower than the one computed for related pellet sample.

Both datasets obtained by pellet and ground pellet are transformed by using the first and the second derivatives, generating four additional datasets.

3.3. Feature Selection and Classification

As described in Subsection 3.1, the NIR spectrophotometer performs spectral analysis over a wide range of frequencies, providing absorption values at each wavelength. As a consequence, each sample is represented by a high-dimensional vector composed of 261 elements (called features). The presence of a large number of features is a well-known problem, usually referred to as *curse of dimensionality*, and negatively impacts on the accuracy of classification. For this reason, we use a feature selection technique to reduce the number of features, by selecting only the most significant ones for

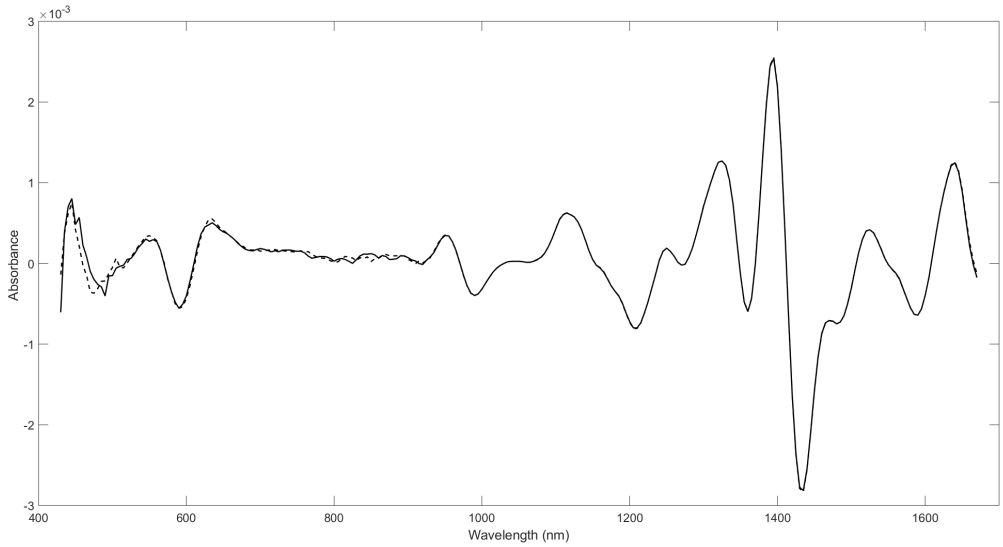

Figure 4: Second derivative of two spectra (dotted and solid lines) obtained repeating the NIR analysis on the same randomly selected ground pellet sample.

the considered problem. The proposed technique is based on the use of four Decision Trees (DTs), each of which is trained on the same dataset using a different splitting criterium, namely Gini index, information gain, accuracy and gain ratio. The technique is an extension of the approach proposed in (Peng et al., 2002) and (Sugumaran et al., 2007), where a single DT is used for feature selection. Since in a DT the best splits are performed early while growing the decision tree, we select only features that appear within the top levels of all induced DTs; the other features can be discarded since they have poor discriminating capability. For an evaluation of the optimal number of features for the considered problem we refer to Section 4.2.

The selected features are used to build a classification model and to predict pellet quality. To this end, we propose two classification approaches, namely automatic and semi-automatic approach. The former works at spectrum level, that is for each sample only one spectrum is collected and the class of the spectrum is assigned to the sample. The latter works at sample level. Given a set of spectra related to the same sample, if spectra are evenly classified then the same class is assigned to the sample; otherwise traditional laboratory analysis is required. In both approaches, the quality class of a pellet batch is given by majority voting of samples classification.

No. of samples		70
No. of spectra		140
No. of features		261
Class distribution	A	48
	B	10
	OOS	12

Table 1: Dataset characteristics.

The use of information from different spectra and, when needed, the use of laboratory analysis make semi-automatic approach more accurate. However, semi-automatic approach is slower than automatic one. Furthermore, automatic approach is advantageous (as well as necessary) when it is difficult to take more than one spectrum from a sample. This is the case, for instance, of pellet unloaded from a ship by a conveyor belt, in which a fixed NIR spectrophotometer can be installed in order to perform real-time analysis. In such case, pellet continuously flows under the NIR device and hence it is not possible to repeat a scan on the same sample.

4. Experiments

This section is devoted to discuss the results of the application of the proposed methodology to real-world data, by training several machine learning algorithms and evaluating the impact of transformations used to reduce both the scattering effect and the number of features.

4.1. Experimental setup

The proposed methodology is evaluated by analyzing 70 pellet samples, which different Italian power plants sent to our laboratory for quality analysis. The sampling period refers to March-May 2017 and February-May 2018. For each pellet sample, two spectra are extracted. Hence, both pellet dataset (P) and ground pellet dataset (GP) have 140 spectra and 261 features. Table 1 provides an overview of datasets characteristics and class distribution.

In next experiments we consider six datasets generated applying preprocessing techniques described in Subsection 3.2: pellet with no transformation (P), pellet using first derivative (P-FD), pellet using second derivative (P-SD), ground pellet (GP), ground pellet using first derivative (GP-FD), and

Algorithm	Parameters
SVM	kernel={linear, rbf(gamma=1), polynomial(degree=2)}, max_iterations=100000
DT	splitting_metric={gini index, information gain, gain ratio}, max_depth=20, min_leaf_size=2
RF	num_trees=100, splitting_metric=information gain
MLP	activaction_function=Maxout, epochs=2000, hidden_layers=2, units_per_layer=50
LDA	solver=least squares
NB	Laplace_correction=true
k-NN	k$\in [3, 10]$

Table 2: List of algorithms and parameters setting.

ground pellet using second derivative (GP-SD). For each dataset, we apply seven classification algorithms, namely Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), MultiLayer Perceptron (MLP), Linear Discriminant Analysis (LDA), Naïve Bayes (NB) and k-Nearest Neighbor (k-NN). Parameters we use to set up algorithms are shown in Table 2. For what concerns k-NN, we tested several values for k (i.e., from K=3 to K=10) but we only report the performance of the best classifier.

We compare classification algorithms by means of two metrics: classification accuracy and F_1 score. Let x_{ij} be the number of data belonging to j-th class which have been classified as i-th class. Let C be the number of classes and N be the total number of data. The accuracy achieved by a classifier is computed as:

$$accuracy = \frac{1}{N} \sum_{i=1}^{C} x_{ii} \tag{1}$$

Precision and recall of i-th class are determined as follows:

$$precision_i = \frac{x_{ii}}{\sum_{j=1}^{C} x_{ij}} \tag{2}$$

$$recall_i = \frac{x_{ii}}{\sum_{j=1}^{C} x_{ji}} \tag{3}$$

Dataset	Best classifier	Accuracy	$\mathbf{F}_1$
P-SD	NB	0.82	0.75
P-FD	RF	0.82	0.68
GP-FD	RF	0.81	0.68
P	MLP	0.79	0.63
GP-SD	SVM (linear kernel)	0.81	0.59
GP	SVM (linear kernel)	0.74	0.46

Table 3: Performance of the best classifier for each dataset (without feature selection), ordered by F_1 score.

F_1 score of i-th class is equal to:

$$F_{1i} = 2 \cdot \frac{precision_i \cdot recall_i}{precision_i + recall_i} \tag{4}$$

Therefore, the F_1 score achieved by a classification model is defined as the average of F_{1i}:

$$F_1 = \frac{1}{C} \sum_{i=1}^{C} F_{1i} \tag{5}$$

The F_1 score is considered since in presence of imbalanced datasets (as the one used in the experiments) it gives more precise information about the classification performance. In order to make the comparison significant, we compute the metrics using the leave-one-out cross validation technique. Since for each sample we have two spectra (which, although not identical, have very similar values), at each iteration of the leave-one-out the test set is composed by the two spectra of the same sample. In this way, we guarantee independence between training and test set. Hence, for each classification algorithm, we define 70 classification models, compute metrics and return average values.

4.2. Evaluation of preprocessing techniques

As first goal of our experiments, we evaluate the impact of transformations used to reduce the scattering effects. To this end, we first analyse results without applying feature selection technique. Table 3 shows the best result in terms of accuracy and F_1 score for each dataset.

The differences between the various configurations in terms of accuracy are rather small, but the highest F_1 values are obtained when analyzing pellet

samples and pre-treating the data with the second derivative. In such case, the Naïve Bayes classification algorithm shows the best performance, with 0.82 of accuracy and 0.75 of F_1.

It is to be noted that each transformation of the spectrum improves performance. Indeed, results obtained using first and second derivatives (P-FD, P-SD, GP-FD and GP-SD) are better than those on corresponding original datasets (P and GP). Furthermore, unlike what one could expect, the grinding process combined with spectral analysis does not contribute to improving pellet classification. Indeed, the physical transformation worsens the results obtained on the original dataset. All three datasets representing ground pellet samples (GP, GP-FD, GP-SD) have less performance than the corresponding datasets with pellet samples (P, P-FD, P-SD). From the point of view of the implementation of the quality control process, this result represents an advantage: as a matter of fact, computing derivatives is a mathematical operation that, unlike grinding, is not time-consuming and does not require additional equipment.

4.3. Feature selection

As second goal of the experimental evaluation, we evaluate the impact of the feature selection technique proposed in 3.3. First, in order to choose the right number of features to be used, we trained all the classifiers several times, increasing the number of features. In details, at iteration i, we selected only features used in the top i levels of the decision trees (as described in Subsection 3.3). Figure 5 shows accuracy achieved by all classification algorithms on the P-SD dataset. In all cases, as the number of levels (and consequently the number of selected features) grow, the accuracy tends to stabilize. The only exception is SVM with rbf kernel in which performance deteriorates. However, this algorithm shows the worst accuracy. Results in Figure 5 are almost constant from the sixth level of the decision trees, corresponding to 21 features. For this reason, in the following experiments we set to 21 the number of features to be selected.

In order to qualitatively evaluate the selected features, in Figure 6 we show the second derivative of the averaged spectra of all pellet samples of each class. Most of the selected features are in the near-infrared region of the electromagnetic spectrum. In particular, the region from 1330 to 1365 nm is assigned to 1st overtone of CH combination bands (CH stretching and CH deformation) of cellulose and hemicellulose compounds (Schwanninger et al.,

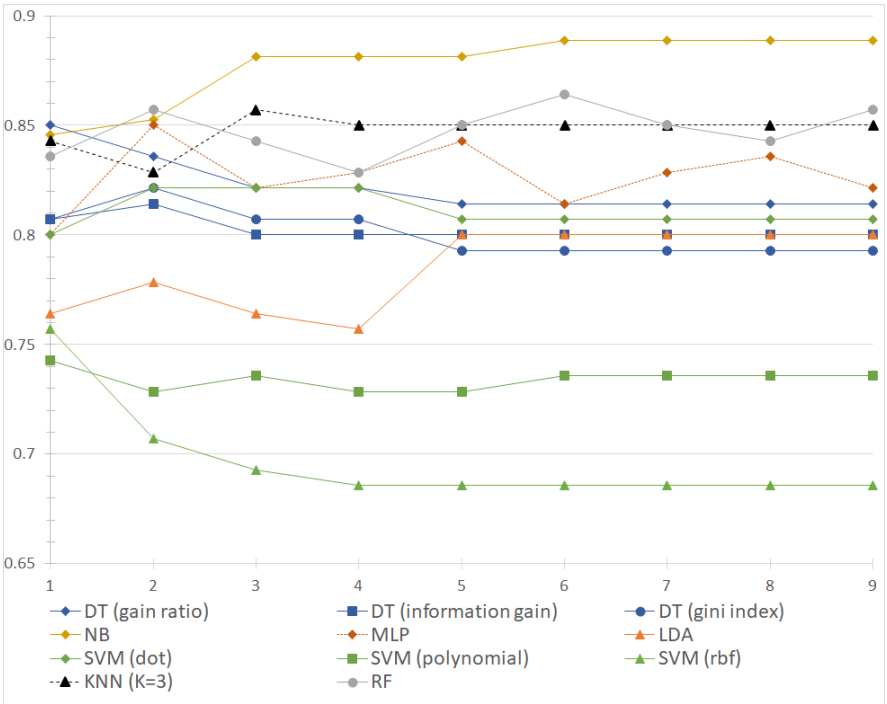

Figure 5: Classification accuracy obtained on P-SD dataset increasing the number of features.

2011; Popescu et al., 2018). The peak at 1605 nm is also related to 1st overtone of CH stretching (Schwanninger et al., 2011). Instead, the 1st and 2nd overtones of CH stretching vibrations in methyl and methylene groups are found in the region from 1100 to 1330 nm (Popescu et al., 2018). In detail, the peak at 1195 nm is assigned to the 2nd overtone asymmetric CH stretching vibration of CH_3 groups in acetyl ester groups (Schwanninger et al., 2011; Popescu et al., 2018). The assignment is in line with bibliography studies (Fagan et al., 2011; Mancini et al., 2018). It is noteworthy that in Literature the presence of CH containing compounds is recognized to be indirectly related to the ash content(Gillespie et al., 2015; Lestander and Rhén, 2005). Hence, having selected features mainly assigned to CH compounds confirms their suitability for predicting pellet quality classes. Other features are related to the visible part of the spectrum. In particular, relevant differences in the averaged spectra of the three quality classes of pellet samples can be seen at peak 435, 665, 670 and 680 nm. Lastly, it is important to note that not all the spectral differences are selected by our features since some of them

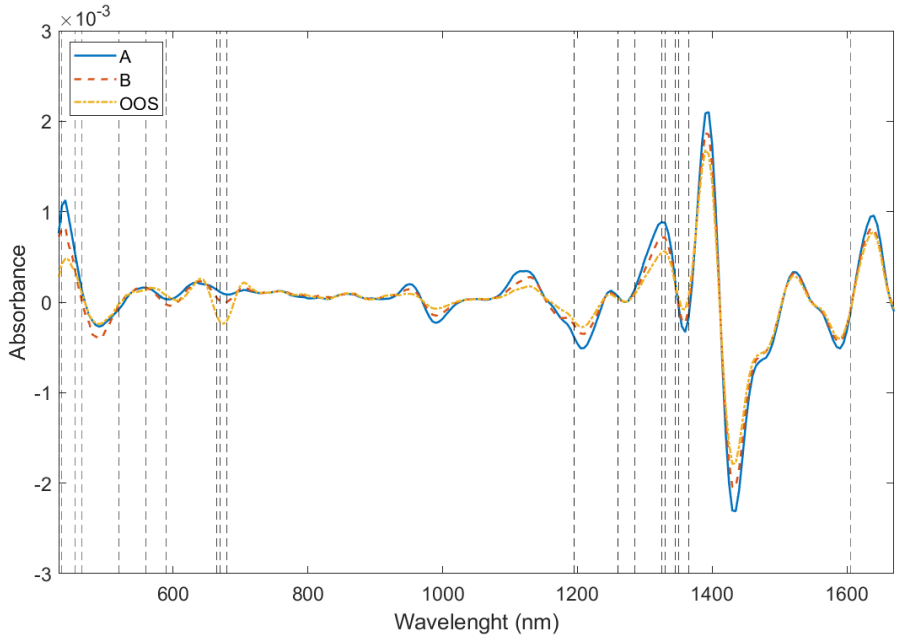

Figure 6: Second derivative averaged spectra of pellet samples divided by A, B and OOS quality classes. The selected features are marked with dotted lines.

are probably not relevant for ash content prediction but could be useful for other quality parameters. For example, the peak at 1435 nm is essential for moisture content prediction, as it is assigned to the OH bonds of water (Schwanninger et al., 2011).

Table 4 reports the best results in terms of accuracy and F_1 score obtained for each dataset when feature selection technique is applied. It is to be noted that the use of feature selection provides a general improvement in terms of classification accuracy and F_1. In particular, the best result is again obtained by the Naïve Bayes algorithm on the pellet dataset preprocessed with second derivative. In this case, we have an increase of 8.53% in accuracy and 12.00% in F_1, if compared with P-SD without feature selection. Analyzing results achieved by all eleven classification models, it turns out that results obtained on P-SD dataset are always better than those obtained on P dataset, in 9 out of 10 models are better than results obtained on GP dataset, and in 8 out 10 cases are better than those achieved on GP-FD and GF-SD datasets. Results on P-SD and P-FD are similar in terms of accuracy (P-SD wins F-PD 6 times out of 10), but P-SD dataset performs better in terms of F_1 score. Hence,

Dataset	Best classifier	Accuracy	$\mathbf{F}_1$
P-SD	NB	0.89	0.84
P-FD	RF	0.86	0.75
GP-FD	MLP	0.85	0.75
GP-SD	MLP	0.83	0.73
GP	MLP	0.74	0.57
P	SVM (linear kernel)	0.79	0.51

Table 4: Performance of the best classifier for each dataset (with feature selection), ordered by F_1 score.

P-SD with feature selection has proved to be the best dataset for predicting pellet quality class.

In order to better appreciate the effect of feature reduction on the classification, in Table 5 are reported the best performance achieved by all classification algorithms on P-SD dataset with and without applying feature selection. It should be noted that, regardless of the classification algorithm, feature selection leads to an improvement in terms of accuracy and F_1. Performance does not change only in the case of SVM with linear kernel. On average, when feature selection is adopted the classification accuracy (F_1 score) increases from 73.42% to 81.25% (from 59.08% to 66.42%). The effect of feature reduction is more evident using LDA, in which accuracy doubles.

4.4. Classification approaches

This subsection is devoted to present the approaches to classify a batch of pellet, namely the automatic and semi-automatic approaches. Basically the former approach classifies each spectrum independently, hence the class of a sample is given by the class of its spectrum. In the latter the classification of a sample is obtained by applying unanimity rule to classification of spectra related to the sample and performing traditional laboratory analysis for uncertain classifications. In both approaches, the quality class of an entire batch corresponds to the majority class of its samples.

In order to analyze performance of the automatic approach, in Table 6 we report the confusion matrix obtained by NB trained on P-SD dataset with feature selection, which achieves the best performance. The three classes have good recall, which means that only few samples for each class are misclassified. For what concerns precision, the values for classes A and OOS are

	No Feature Selection		Feature Selection	
Classifier	**Accuracy**	**F_1**	**Accuracy**	**F_1**
LDA	0.40	0.37	0.80	0.65
DT (gain ratio)	0.74	0.65	0.81	0.71
DT (information gain)	0.74	0.54	0.80	0.65
DT (gini index)	0.71	0.59	0.79	0.66
SVM (linear)	0.82	0.57	0.82	0.57
SVM (rbf)	0.69	0.42	0.70	0.44
SVM (polynomial)	0.71	0.44	0.73	0.46
KNN (K=3)	0.79	0.71	0.85	0.73
NB	0.82	0.75	0.89	0.84
MLP	0.82	0.68	0.85	0.74
Random Forest	0.79	0.69	0.86	0.78

Table 5: Best performance achieved by all classification algorithms on P-SD, without and with feature selection.

high while the value for B class is low: only 17 out of 30 spectra predicted as B actually belong to this class. This could be explained by considering that the B class has intermediate values of ash content and hence same samples could show some similarities with the other two classes. Analyzing misclassified spectra, it is noteworthy that 5.71% (i.e., 8 samples) of spectra are assigned to a higher quality class: the 2.14% of B class are assigned to A class, and 3.57% of OOS class to B class. These are fraudulent behaviors, because it means to consider acceptable for residential use some samples that can be used only in industry (B samples classified as A), and for industrial

	Actual A	**Actual B**	**Actual OOS**	**Precision**
Predicted A	88	3	0	0.97
Predicted B	8	17	5	0.57
Predicted OOS	0	0	19	1
Recall	0.92	0.85	0.79	

Table 6: Confusion matrix related to NB on P-SD with FS.

	Actual A	**Actual B**	**Actual OOS**	**Precision**
Predicted A	43	1	0	0.98
Predicted B	3	8	1	0.67
Predicted OOS	0	0	8	1
Undefined	2	1	3	
Recall	0.93	0.89	0.89	

Table 7: Confusion matrix related to NB on P-SD with FS. Sample classification with undefined class.

use samples that can be used only in special industrial power plants (OOS samples classified as B). On the contrary, the 8 spectra belonging to A class that are labelled as B represent a loss of gain for the manufacturer, as the selling price depends on the quality class.

It is noteworthy that, in the automatic approach, for each of the 70 classifiers (obtained applying leave-one-out), the two spectra representing the same sample are both in the test set, but are classified independently. In semi-automatic approach we take advantage of having replications of spectra to increase the accuracy of prediction. To this end, we propose a reformulation of the classification rule, as follows: if the two spectra are classified as belonging to the same class i, then the sample will be labelled as i; otherwise the classification is uncertain and the sample will be labelled as *undefined*. In this way, some samples (i.e., undefined ones) cannot be automatically classified and require further investigation in a laboratory, which returns a correct classification. The effect of this rule on the confusion matrix shown in Table 6 is reported in Table 7, where each matrix entry represents number of samples instead of spectra.

Now, only the 7.14% of samples are misclassified, returning an accuracy of 0.92 and an average F_1 of 0.89. Furthermore, both precision and recall values are higher than ones reported in Table 6. It is noteworthy that only 2 misclassified samples (i.e., less than 3%) are assigned to higher quality class. Since laboratory analysis is a time-consuming task, unlike the automatic approach, the use of undefined elements is not suitable for real-time analysis. However, it considerably reduces the number of pellet samples needing laboratory analysis (i.e., only the 8.57% of samples). Hence, the semi-automatic

approach is both cheaper and quicker than the traditional approach, which requires the laboratory analysis for all samples.

4.5. Discussion

On the basis of the experiments presented in previous subsections, we can define the best approach for evaluating pellet quality at each phase of the supply chain. It consists of the following steps:

- **NIR Data Gathering**: given a batch of pellet, there are two main ways to acquire samples: using an on-line NIR spectophotometer (which can be easily installed on a conveyor belt) or a hand-held NIR device. In the former case, spectra are continuously acquired, while in the latter two spectra are gathered from each randomly picked samples. It is to be noted that in both cases the acquisition of a spectrum through NIR devices requires a very short time;

- **Preprocessing**: in order to reduce the scattering effect, spectra are pre-processed by computing second derivate;

- **Feature Selection**: data dimensionality is reduced by only selecting predefined relevant features, as described in Subsection 4.3.

- **Classification**: pellet samples are assigned to quality classes by using a pre-trained classification model, according to two approaches:

  - automatic approach: classification is performed at spectrum level, namely the quality class of a sample is the same as that of the corresponding spectrum. This approach is suited for fixed NIR devices (e.g., installed on a conveyor belt);
  - semi-automatic approach: classification is performed at sample level. Given a sample, if the two spectra are evenly classified, then the quality class is assigned to the sample; otherwise a laboratory analysis is required.

  In both cases, the class with the majority of samples is assigned to the batch.

Since NIR data gathering requires a very short time and no physical transformations (e.g., grinding) is needed, the proposed methodology allows

to easily collect data from several samples. This overcomes the problem of sample representativeness, ensuring a more accurate assignment of the entire batch to a quality class. Furthermore, using a hand-held device, a higher number of spectra for each sample can be also collected, allowing for the reduction of measurement uncertainty and the achievement of a more accurate classification of a sample. In this case, in order to reduce the request for laboratory analysis, a majority voting mechanism can be introduced instead of unanimity rule.

The choice between the two classification approaches depends on two main factors: (i) the physical characteristic of the place where acquisitions are performed (e.g., unloading of a ship, warehouse, power plant); (ii) the trade-off between confidence on classification results and time/cost saving. Indeed, the semi-automatic approach provides higher accuracy but it is more expensive and time-consuming than the automatic approach, as it could require some laboratory analysis.

5. Conclusion

The goal of this work is the introduction of a methodology for quick and low-cost classification of pellet quality. To this end, machine learning techniques are used to classify pellet spectra obtained by means of NIR spectroscopy. In details, we propose two different approaches: i) a spectrum-level automatic classification of pellet, which does not require laboratory analysis, and ii) a sample-level semi-automatic approach, which increases the overall accuracy but requires laboratory analysis for uncertainly classified samples. We validate the methodology by performing several experiments on real-world data, training different machine learning algorithms and evaluating the impact of transformations used to reduce both the scattering effect and number of features.

In detail, pellet quality class is predicted with a classification accuracy of 89% using the automatic approach and with an accuracy of 92% using the semi-automatic approach. These results show that machine learning coupled with NIR method is a valid alternative to the traditional laboratory analysis, opening new perspectives for the sector.

Specifically, the speed of analysis and dimensions of the device is a characteristic that makes NIR method suitable in different operational contexts along the supply chain (e.g., ship unloading, pelletizing process, warehouse,

power plant), allowing sector operators to perform a high number of analysis in short time and to have a real time quality control of the product.

We plan to extend the experimental evaluation by using a larger dataset of pellet samples, increasing both the number of collected spectra and the variability of pellet in terms of producers and raw materials. Furthermore, in order to improve the industrial application of the methodology, we plan to analyze the confidence of classification with respect to the number of spectra acquired for each sample.

Acknowledgements

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

Duca, D., Riva, G., Pedretti, E. F., and Toscano, G. (2014). Wood pellet quality with respect to en 14961-2 standard and certifications. *Fuel*, 135:9 – 14.

Espinoza, J., Hodge, G., and Dvorak, W. (2012). The potential use of near infrared spectroscopy to discriminate between different pine species and their hybrids. *Journal of Near Infrared Spectroscopy*, 20:437–447.

Fagan, C. C., Everard, C. D., and McDonnell, K. (2011). Prediction of moisture, calorific value, ash and carbon content of two dedicated bioenergy crops using near-infrared spectroscopy. *Bioresource Technology*, 102(8):5200 – 5206.

Filbakk, T., Skjevrak, G., Hoibo, O., Dibdiakova, J., and Jirjis, R. (2011). The influence of storage and drying methods for scots pine raw material on mechanical pellet properties and production parameters. *Fuel Processing Technology*, 92(5):871 – 878.

García, R., Gil, M., Rubiera, F., and Pevida, C. (2019). Pelletization of wood and alternative residual biomass blends for producing industrial quality pellets. *Fuel*, 251:739 – 753.

Gillespie, G. D., Everard, C. D., and McDonnell, K. P. (2015). Prediction of biomass pellet quality indices using near infrared spectroscopy. *Energy*, 80:582 – 588.

IEA (27 November 2017). Technology roadmap - delivering sustainable bioenergy.

Kusumoputro, B., Faqih, A., and Sutarya, D. (2013). Quality classification of green pellet nuclear fuels using radial basis function neural networks. In *2013 12th International Conference on Machine Learning and Applications*, volume 2, pages 194–198. IEEE.

Lestander, T. A. and Christofer, R. (2005). Multivariate nir spectroscopy models for moisture, ash and calorific content in biofuels using bi-orthogonal partial least squares regression. *Analyst*, 130(8):1182 – 1189.

Lestander, T. A., Finell, M., Samuelsson, R., Arshadi, M., and Thyrel, M. (2012). Industrial scale biofuel pellet production from blends of unbarked softwood and hardwood stems-the effects of raw material composition and moisture content on pellet quality. *Fuel Processing Technology*, 95:73 – 77.

Lestander, T. A., Johnsson, B., and Grothage, M. (2009). Nir techniques create added values for the pellet and biofuel industry. *Bioresource Technology*, 100(4):1589 – 1594.

Lestander, T. A. and Rhén, C. (2005). Multivariate nir spectroscopy models for moisture, ash and calorific content in biofuels using bi-orthogonal partial least squares regression. *Analyst*, 130:1182–1189.

Magelli, F., Boucher, K., Bi, H. S., Melin, S., and Bonoli, A. (2009). An environmental impact assessment of exported wood pellets from canada to europe. *Biomass and Bioenergy*, 33(3):434 – 441.

Mancini, M., Rinnan, A., Pizzi, A., and Toscano, G. (2018). Prediction of gross calorific value and ash content of woodchip samples by means of ft-nir spectroscopy. *Fuel Processing Technology*, 169:77 – 83.

Maranan, C. M. and Laborie, M. G. (2008). Rapid prediction of the chemical traits of hybrid poplar with near infrared spectroscopy. *Journal of Biobased Materials and Bioenergy*, 2:57– 63.

Mola-Yudego, B., Selkimaki, M., and Gonzalez-Olabarria, J. R. (2014). Spatial analysis of the wood pellet production for energy in europe. *Renewable Energy*, 63:76 – 83.

Monti, A., Di Virgilio, N., and Venturi, G. (2008). Mineral composition and ash content of six major energy crops. *Biomass and Bioenergy*, 32(3):216 – 223.

Nielsen, N., Gardner, D., Poulsen, T., and Felby, C. (2009). Importance of temperature, moisture content, and species for the conversion process of wood residues into fuel pellets. *Wood and Fiber Science*, 41:414–425.

Pasquini, C. (2003). Near infrared spectroscopy: fundamentals, practical aspects and analytical applications. *Journal of the Brazilian Chemical Society*, 14.

Peng, Y., Flach, P., Brazdil, P., and Soares, C. (2002). Decision tree-based data characterization for meta-learning. In *ECML/PKDD-2002 Workshop IDDM-2002*, pages 188–195.

Popescu, C.-M., Navi, P., Peña, M. I. P., and Popescu, M.-C. (2018). Structural changes of wood during hydro-thermal and thermal treatments evaluated through nir spectroscopy and principal component analysis. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 191:405 – 412.

Quddus, M. A., Hossain, N. U. I., Mohammad, M., Jaradat, R. M., and Roni, M. S. (2017). Sustainable network design for multi-purpose pellet processing depots under biomass supply uncertainty. *Computers & Industrial Engineering*, 110:462 – 483.

Rinnan, A. (2014). Pre-processing in vibrational spectroscopy - when, why and how. *Analytical Methods*, 6:7124–7129.

Rinnan, A., van den Berg, F., and Balling Engelsen, S. (2009). Review of the most common pre-processing techniques for near-infrared spectra. *TrAC Trends in Analytical Chemistry*, 28(10):1201 – 1222.

Sandak, A., Sandak, J., and Negri, M. (2011). Relationship between near-infrared (nir) spectra and the geographical provenance of timber. *Wood Science and Technology*, 45(1):35–48.

Santoni, I., Callone, E., Sandak, A., Sandak, J., and Diré, S. (2015). Solid state nmr and ir characterization of wood polymer structure in relation to tree provenance. *Carbohydrate Polymers*, 117:710 – 721.

Savitzky, A. and Golay, M. J. E. (1964). Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry*, 38(8):1627–1639.

Schwanninger, M., Rodrigues, J. C., and Fackler, K. (2011). A review of band assignments in near infrared spectra of wood and wood components. *Journal of Near Infrared Spectroscopy*, 19(5):287–308.

Selkimaki, M., Mola-Yudego, B., Roser, D., Prinz, R., and Sikanen, L. (2010). Present and future trends in pellet markets, raw materials, and supply logistics in sweden and finland. *Renewable and Sustainable Energy Reviews*, 14(9):3068 – 3075.

Sugumaran, V., Muralidharan, V., and Ramachandran, K. (2007). Feature selection using decision tree and classification through proximal support vector machine for fault diagnostics of roller bearing. *Mechanical Systems and Signal Processing*, 21(2):930 – 942.

Sun, L. and Niquidet, K. (2017). Elasticity of import demand for wood pellets by the european union. *Forest Policy and Economics*, 81:83 – 87.

Sutarya, D. and Kusumoputro, B. (2011). Quality classification of uranium dioxide pellets for pwr reactor using anfis. In *TENCON 2011-2011 IEEE Region 10 Conference*, pages 118–123. IEEE.

Toscano, G., Duca, D., Foppa Pedretti, E., Pizzi, A., Rossini, G., Mengarelli, C., and Mancini, M. (2016). Investigation of woodchip quality: Relationship between the most important chemical and physical parameters. *Energy*, 106:38 – 44.

Toscano, G., Riva, G., Foppa Pedretti, E., Corinaldesi, F., C., M., and D., D. (2013). Investigation on wood pellet quality and relationship between ash content and the most important chemical elements. *Biomass and Bioenergy*, 56:317 – 322.