Project Acronym:	PhasmaFOOD
Grant Agreement number:	732541 (H2020-ICT-2016-1 - RIA)
Project Full Title:	Portable photonic miniaturised smart system for on-the-spot food quality sensing

DELIVERABLE

Deliverable Number	**D3.3**
Deliverable Name	**Detection algorithms**
Dissemination level	Public
Type of Document	Research outcomes
Contractual date of delivery	June 30th 2018 (M18)
Deliverable Leader	DLO
Status & version	Final – v5.0
WP / Task responsible	WP 3 / all tasks
Keywords:	Chemometrics, multivariate analysis, data fusion
Abstract (few lines):	This deliverable reports on the multivariate analysis and data fusion strategy developed (together 'Detection algorithms') to predict the contamination, spoilage and authenticity issues for the spectral data of the PhasmaFOOD micro-sensors. For each use case, a different useapproach is proposed to be implemented in the PhasmaFOOD embedded software and/or cloud computing service. This is necessary, as each use case deals with either targeted (e.g. mycotoxins) or untargeted (e.g. fraud detection) assessment of food samples in either a classification or regression model. The integration of the developed detection algorithms in the PhasmaFOOD software platform is touched upon as a final issue.

This document is part of a project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 732541. It is the property of the PhasmaFOOD consortium and shall not be distributed or reproduced without the formal approval of the PhasmaFOOD Management Committee.

Deliverable Leader:	Martin Alewijn, Yannick Weesepoel (DLO)
Contributors:	Eugenio Martinelli (UTOV), Annamaria Gerardino (CNR), Francesca Romana Bertani (CNR), Panagiotis Tsakanikas (AUA), Milenko Tosic (VLF)
Reviewers:	CNR and WINGS
Approved by:	INTRASOFT

Executive Summary

PhasmaFOOD touches upon three use cases in which the effectiveness of the sensor array is addressed for versatile food safety and food authenticity challenges. Vital to processing and making sense out of the raw PhasmaFOOD sensor signals is the application of multivariate statistics or chemometrics in order to classify or determine the concentration of a compound of interest in the food. Since the PhasmaFOOD array contains three spectral sensors delivering unidimensional data and one micro-camera delivering bi-dimensional data, strategies for the effective combining of these data-streams is vital for successful application of the device. In this deliverable, the chemometric strategy and data-fusion strategy is elaborated upon individually, as each use case has specific requirements. For the aflatoxin case, the specific chemical of interest is known, whilst for the food fraud case, any chemical of interest needs to be considered. The food spoilage case deals more with regression modelling than with classification modelling and therefore it also needs a unique approach. In this deliverable, the integration of the bi-dimensional data from the micro-camera is not elaborated upon. Finally, we discuss in conjunction with D4.1 the integration of the algorithms and the decision making process in the cloud of the PhasmaFOOD sensor array.

Project Title: PhasmaFOOD
Contract No. 732541
Project Coordinator: INTRASOFT International S.A.

Document History			
Version	**Date**	**Contributor(s)**	**Description**
1.0	30/04/2018	DLO, CNR, UTOV, AUA	First draft
2.0	08/06/2018	DLO, CNR, UTOV, AUA, VLF	Concept for consortium review
3.0	18/06/2018	DLO, UTOV, AUA, FUB	Concept for consortium review
4.0	25/06/2018	DLO, CNR, WINGS	Reviewed document for CL
5.0	29/06/2018	INTRA	Final document

Table of Contents

Project Title: PhasmaFOOD
Contract No. 732541
Project Coordinator: INTRASOFT International S.A.

Table of Figures

List of Tables

Definitions, Acronyms and Abbreviations

Acronym	Title
DA/ML	Data analysis and Machine Learning
FLU	Fluorescence
FTIR	Fourier Transformation Infrared
IaaS	Infrastructure as a Service
k-NN	k-Nearest Neighbors
LDA	Linear Discriminant Analysis
LED	Light Emitting Diode
MSI	Multi-Spectral Imaging
NIR	Near-Infrared
OCC	One-Class Classification
OSVM	One-Class Support Vector Machine
PCA	Principal Component Analysis
PLS	Partial Least Squares
PLSR	Partial Least Squares Regression
RF	Random Forest
RNV	Robust Normal Variate
SIMCA	Soft Independent Modelling of Class Analogy
SMP	Skimmed Milk Powder
SNV	Standard Normal Variate
VIS	Visible

1 General information

1.1 Scope

This deliverable on detection algorithms describes the (mathematical) processes needed to transform a spectroscopic raw signal into a result, i.e. a qualitative or quantitative assessment of the sample. These processes include steps that are based on a database of previously measured (reference) samples, so the building of such a database and the building of a mathematical classification/regression model based on the samples in the database is an integral part of this process. Furthermore, integration of the detection algorithms in the PhasmaFOOD cloud is elaborated upon in conjunction with D4.1. Data-processing strategies concerning the micro-camera will be elaborated upon in D3.5, integrated device protocols and algorithms, by M27 (March 2019).

1.2 Data preprocessing

The first step in this process is the data pretreatment, the process between the signal in the (electronic) format, generated by the analytical device and the form in which it is fed to the chemometric algorithm. After transfer and possible format conversion from the device into the cloud or a computer environment, a vast number of options exist, but there are two main reasons to perform data pretreatment.

- One very common task in (handheld) spectroscopy is to adjust the intensity of the signal. Reference measurements may be used to compensate the signal for fluctuations in the light source, but also light scatter effects of the sample, fluctuations in the angle in the light–sample–detector trajectory and/or stray light from the environment may influence the raw signal and should be corrected for.
- One other reason to perform data pretreatment is to reduce noise or enhance certain signal features. Examples are variable selection, spectrum derivatives or wavelet transforms. As with any pretreatment, the conditions of the raw data and the intended application determine the need and applicability of the different forms of data pretreatment.

1.3 Chemometric algorithms

A chemometric algorithm is the (mathematical) process of translating the (preprocessed) data into a result. The result may be a continuous number (concentration, time) and then we develop a regression algorithm, or the result may be categorical (present/absent, safe/unsafe), for which a classification algorithm is needed. Both for regression and classification, a large number of different algorithms is available, each with different characteristics. There is no established way

of selecting 'the best' chemometric algorithm for a specific case. As long as the algorithm is well described and is properly validated before practical use, any algorithm can be used.

1.4 Data fusion strategies

One of the unique features of the PhasmaFOOD project is that samples are analyzed (near)-simultaneously by three different techniques, visible (VIS), fluorescence (FLU) and near-infrared (NIR). These techniques provide distinctly different pieces of physical/chemical information about the sample. Depending on the specific case, one or more of the three techniques contain information that leads to the desired result. Especially, if multiple sensors carry information on different aspects of the intended result, a combination of their results will lead to better results. Data fusion is the process of combining results from multiple techniques.

The scientific literature does not provide one single definition or process to perform data fusion and – much like many other aspects of chemometrics – there are many ways to do so [2, 3]. When working with results from three different sensors on the same sample, there are roughly three ways of combining the data:

1. Combine the (preprocessed) data from all sensors and feed this into the model
2. Extract a limited number of features – individual points in the spectrum or combinations of several points generated by a certain algorithm – from each sensor based on a set of reference samples with known results, and combine those features as input for an overall chemometric model
3. Develop chemometric models to generate results based on all three sensors separately and devise rules to combine the results to the final result.

All these data fusion strategies can and should be fine-tuned depending on the case, and all have their strengths and weaknesses. The data fusion strategy should be validated prior to practical use.

1.5 Integration of smart algorithms in the cloud interface

As a final step, the procedure developed in the steps above, typically by the different laboratories involved in this project, should be translated into a system that yields similar results without any human guidance. The first software prototype, including data analysis in the cloud, has been described in D4.1.

1.6 Calibration transfer and database exploitation

In this initial phase of the project, the experimental measurements have been performed considering the single sensing elements (VIS, FLU, NIR) using different experimental set-ups and different electronic devices (LED, electronic components, etc.). As a consequence, when the final prototype will be ready, an additional data analysis step, often called calibration transfer, will be

devoted to studying how the database already collected could be exploited with the new PhasmaFOOD device. To reach this goal an *ad-hoc* procedure will be designed measuring the same samples with the old and new PhasmaFOOD prototype.

2 Use case 1 – Detection of mycotoxins in grains and nuts

2.1 Chemometric strategy

The data collected in the frame of use case 1 are related to the sub-use cases on grained almonds and maize. In the case of almonds, the samples have been artificially contaminated while in the case of maize both artificially and naturally contaminated samples have been measured. A table with all the measurement sessions can be found in D3.2 (Table 2) together with the first results of the data analysis.

2.1.1 Preprocessing strategy

As regards use case 1 on detection of mycotoxins, the acquired data comes from the instrument in .csv format. The data treatment is similar in the three kinds of spectroscopy, VIS, FLU and NIR. At the beginning of the measurement session, a dark reference (D), a measurement made with all illumination off and instrument shielded from ambient light, is acquired together with a reference spectrum (Ref), performed by using a reference standard. The spectrum acquired during the measurement (Sp) are then pretreated following formula (1):

$$Sp1 = \frac{Sp-D}{Ref-D} \quad (1)$$

where Sp1 is the spectrum after the pretreatment.

The difference between the VIS, FLU and NIR spectroscopy data is in the used reference spectra because each wavelength range has its own reference standard.

2.1.2 Regression model & classification model

After the spectrum normalization, the acquired spectra are then used as input for the following data processing step. This part of the analysis can be devoted to the estimation of the level of contamination or to the classification of the samples as aflatoxin contaminated or as safe. In the first case, a Partial Least Square (PLS) techniques are used as regression model while a linear classification is used for the classification task.

The general idea of PLS is, given an X input data matrix and Y variable vector to estimate, to try to extract the latent factors of X and Y, accounting for as much of the manifest factor variation

Project Title: PhasmaFOOD
Contract No. 732541
Project Coordinator: INTRASOFT International S.A.

as possible while modelling the responses well. We used this technique for compensating for the limited amount of measurement data and for its intrinsic capability to select the most informative features of the spectrum. More details about this technique can be found in D4.1 and [4, 5].

The data analysis technique used in the preliminary phase for the classification task is Fisher discriminant analysis, often called simply, Linear Discriminant Analysis (LDA). The goal of the classifier is to project the original dataset onto a lower-dimensional space with good class-separability in order to avoid overfitting in particular with not so large dataset and also reduce computational costs. More details about the LDA are described in D4.1 and [6].

The choice of the PLS and LDA is only preliminary and a more accurate choice of the models for regression and classification will be considered when we will collect data with the PhasmaFOOD prototype. In that occasion, more complex models such as neural network, support vector machine, deep learning architecture etc. could be also considered to reach the project goals [6].

2.2 Data fusion strategy

To improve the performances of the PhasmaFOOD device for this case study, we will investigate two additional strategies based on static and dynamic data fusion algorithm. In both cases, we will investigate the possibility to use besides the FLU spectrum also the VIS and NIR data to identify and/or to estimate the sample contamination. In the static data fusion approach, the model will be created using the initial training phase and it will not be changed upon future usage. In the dynamic data fusion, an online feature selection of all relevant PhasmaFOOD data will be performed for each new sample comparing the reliability of every single feature of the sample data (Fig. 1).

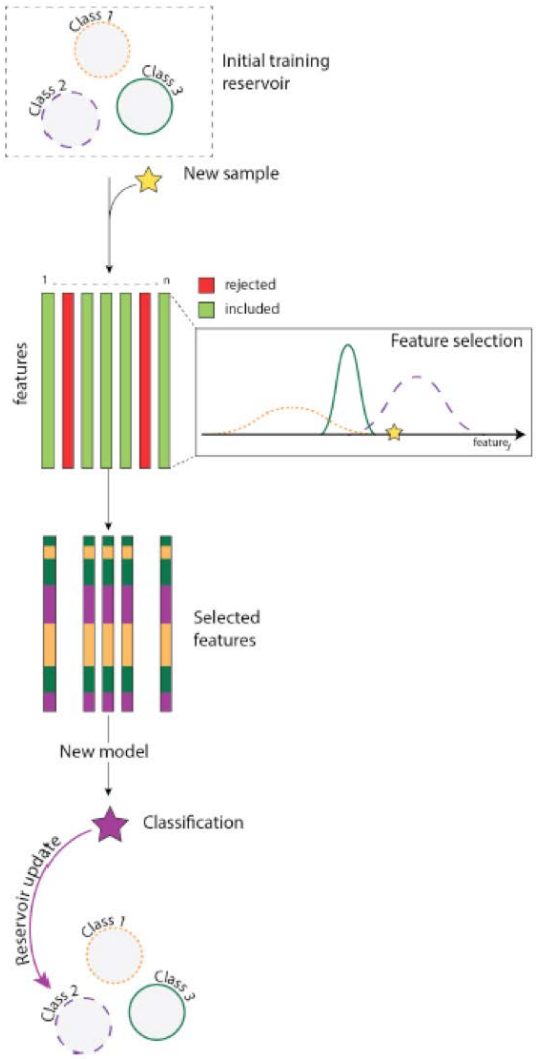

Figure 1 – a sketch of the dynamic feature selection strategies (more details in *[1]*)

The selected feature will be used to train a new model tailored for the specific sample under test. This procedure will be repeated for each new sample. Although this last approach has obtained promising results in the literature, showing also the capability to counteract possible feature outliers, it requires a continuous training procedure and for this reason could not be feasible for the PhasmaFOOD prototype. More details about the dynamic feature selection are found in [1].

3 Use case 2 – Detection of an early sign of spoilage, spoilage and shelf-life estimation in fruits, vegetables, meat and fish

3.1 Meat

3.1.1 Chemometric strategy

The analysis of data derived from the minced meat (pork) spoilage experiments (see D3.2) is in progress and will be reported in full in the further course of WP3. The data analyzed up to this point are corresponding to aerobic storage of minced pork and generated from the multi-spectral imaging (MSI) and Fourier transform infrared (FTIR) spectroscopy sensors. The collected MSI and FTIR data were subjected to pre-processing, *i.e.* smoothing based on standard normal variate (SNV) [7] transformation and the Savitzky-Golay algorithm [8], respectively. Partial least squares regression (PLSR) [9] was used to establish the correlation between imaging/spectral data and microbial counts, with the former constituting the input and the latter the output variables in the PLSR models. The models were calibrated and validated with the data collected from the studied isothermal (170 samples) and dynamic temperature (58 samples) conditions, respectively. In addition, the developed workflow described next in Section 3.3 will also be applied to meat data in the context of the search for a unified data analysis workflow.

3.1.2 Data fusion strategy

At the time of writing the present report, no fusion strategies have been applied to the meat sub-use case. The data fusion research has been initiated in the case of vegetables, which was the first dataset where at least two PhasmaFOOD sensors were available and respective data have been collected (please refer to Section 3.3.2). A recently completed experiment of minced pork stored under modified atmosphere packaging has produced data from at least two PhasmaFOOD sensors and fusion strategies will be developed in the further course of WP3.

3.2 Fish

3.2.1 Chemometric strategy

From the first fish fillets experiment (pilot data acquisition) the collected data for all available, at the time, sensors (i.e. NIR, FTIR and MSI) have been analyzed. Specifically, with regard to the FTIR data related to the skin side of the tested fish samples, the second derivative of the acquired

spectra in the ranges of 3100-2700 and 1800-900 cm^{-1} was calculated, using the second derivative Savitzky-Golay numerical algorithm[8] with a second-order polynomial and a 9-point window. A PLSR model [9] was then trained and developed with the number of significant latent variables being determined based on the results of leave-one-out cross-validation. A further feature selection step was performed using Martens Uncertainty Test [10]. In the case of the MSI data, PLSR was applied using the 18 raw (i.e. no pre-processing step) mean and standard deviation values of the reflectance spectra, both in the case of skin and flesh surfaces of the fish samples.

For all the data categories described above, model training (calibration) was based on the spectral and microbiological data derived from the fish storage experiments at 0 and 8°C (n=96), whereas model testing (prediction) was performed using external data sets obtained during storage of fish samples at the intermediate temperature of 4°C (n=62). In the case of the MSI data, in addition to the aforementioned data partitioning scheme, an alternative training and testing dataset structure were also applied in order to enhance the efficiency of the prediction models (please refer to section 3.3). In the latter scheme, the testing dataset was derived by random sampling of the full data set (~20% of the total sample size using a uniform random generator), consisting of 29 and 30 samples for flesh and skin samples, respectively.

At this point, we must state that the experiment and data analysis presented regarding fish samples was the first of the project. Thus, several issues were, at that time, raised for the NIR sensor, and in this context, the acquired NIR data were excluded from analysis as not reliable. On the other hand, the experience was gained, the issues have been solved, and the following experiments (i.e. storage of fish under modified atmosphere packaging conditions) were performed without any particular technical issues. The latter experiments were recently completed and, thus, the generated data will be analyzed in the further course of WP3.

3.2.2 Data fusion strategy

At the time of writing this report, no fusion strategies have been applied to the fish sub-use case. The data fusion research has been initiated in the case of vegetables, which was the first dataset where at least two PhasmaFOOD sensors were available and respective data have been collected (please refer to Section 3.3.2). A recently completed experiment of fish stored under modified atmosphere packaging has produced data from at least two PhasmaFOOD sensors and fusion strategies will be developed in the further course of WP3.

3.3 Fruit and vegetables

3.3.1 Chemometric strategy

Herein, a detailed description of the data analysis workflow followed in the case of the rocket salad sample dataset (see D3.2) is provided. The developed workflow will be applied to all fruit and vegetable data acquired, i.e. pineapple and baby spinach. The developed pipeline is common for all sensors and consists of 1) feature selection (specific wavelengths) on the basis of random forest (RF) regression ensemble [11] followed by 2) PLSR for microbial contamination and shelf life estimation using the features (wavelengths/wavenumbers) selected by the RF ensemble. Prior to feature selection, spectra were normalized under the SNV normalization scheme [7] and more specifically its robust version, robust normal variate (RNV) [12] (equation 2). This choice is justified since apart from correlated information and multiplicative noise reduction, the robust version of SNV also gives more realistic results (i.e. without artefacts), leading to improved analysis downstream.

$$s_i^{snv} = \frac{s_i - \text{median}(S)}{mad(S)} \quad (2)$$

where S is the ensemble of all spectra, and s_i and s_i^{snv} the i^{th} and the corresponding normalized spectra, and *mad* the median absolute deviation (of S), respectively.

Next, a feature selection step was introduced. This is a critical information extraction step, not only in this specific case but in every regression/classification problems where a small amount of samples in high dimensional space (i.e. large number of variables) is to be exploited. So, in order not to fall into overfitting, variable/feature set is decreased in a way that meaningful features are preserved while irrelevant (to the prediction of microbial contamination and shelf-life) and redundant ones are excluded. RFs were selected to be employed for regression [11], which is an ensemble learning method for regression that constructs a multitude of decision trees at training time while provides as output the mean prediction (regression) of the individual trees. One justification for this selection is that RFs correct decision trees' "habit" of overfitting to their training set, something which is needed here as stated earlier. In addition, RFs under boosting [13] scheme was employed for learning which has the advantage over averaging [11] or bagging [14] since boosting algorithms consist of iteratively learning weak classifiers with respect to a distribution and adding them to a final strong classifier. Having in mind the aforementioned advantages of RFs, their application is expected to output concrete and representative (yet not redundant) features from each sensor, in terms of discriminating the inherent microbial burden.

In detail, the regression tree ensemble was trained using LSBoost (gradient boosting strategy applied for least squares) [15] and 100 learning cycles, with all constant temperatures (i.e. 4, 8 and 12°C) being used as training set and set of samples stored at dynamic temperature conditions being utilized as external test set.

The feature set selected by RFs was fed into a PLS regression scheme [9], a popular method in food quality applications [16, 17]. The training process was again performed on the dataset of samples stored in constant temperatures as previously and 10-fold cross-validation and 10 Monte-Carlo repartitions of the data were performed. Finally, it should be noted that the aforementioned procedure was applied not only for microbial contamination but also for time-on-shelf estimation and prediction. Time-on-shelf term is defined as the time of the product after its production, i.e. the storage time.

3.3.2 Data fusion strategy

Fusion of data derived from T3.2 has been initiated in the sub-use case where at least two micro-sensors were available and used in experiments. With regard to the rocket sub-use case, the acquired NIR and VIS data were subjected to RNV normalization (as defined in the previous section). Then, a vector for each sample containing both sensors' features was created, and RF-based feature extraction was employed. PLSR was finally performed with the selected features (i.e. wavelengths), correlating the spectral data to the microbiological burden. Based on the results obtained so far, data fusion did not improve the results of each sensor alone. So, a possible workaround would be that data will be fused by dot multiplication. Furthermore, additional data fusion strategies will be investigated and composites of individual (sensor-specific) models will be evaluated for the purpose of improving prediction efficiency.

4 Use case 3 – Detection of food fraud: adulteration of milk powder, meat, alcoholic beverages and edible oils

In the food fraud detection case, so far only the meat and milk powder sub-use cases have been studied more extensively using the FLU, VIS and NIR spectral data. The data derived from the recently completed minced meat adulteration experiments have not been analyzed yet and will be reported in the further course of WP3. Hence, the detection algorithms described in this report for use case 3 will only refer to the milk powder sub-use case.

4.1 Chemometric strategy

As reported in D3.2, 32 authentic skimmed milk powders (SMP) have been measured, as well as 72 different mixes of several SMPs with several bulk and chemical adulterants: For the non-hazardous fillers, whey protein isolates (6 commercial brands), plant protein isolates (5 commercial brands), pea protein isolates (4 commercial brands), soy protein isolates (3 commercial brands), buttermilk powder (reference standard), starch, maltodextrin, glucose, lactose and fructose were considered. Due to time constraints, the pure milk samples were measured in triplicate, but the adulterants were measured only once, and need to be repeated as soon as the sensors become available again.

The first chemometric step in the process is the evaluation of the quality of the (spectral) data. This was done separately for each sensor type and consisted of reviewing the data using principal component analysis (PCA) on the raw SNV-scaled data. We observed good spectral quality for all three sensors, where even the extreme wavelengths do not suffer from extensive noise, unlike some commercially available scanners. However, significant day-to-day bias was observed, which will be corrected for using appropriate reference measurements in future analyses. For the current data set where reference measurements were not recorded, we estimated the day-to-day bias by taking the average of at least eight spectral differences for samples that were repeated on both days. Subsequently, we used the simple addition of this average spectral difference to one single day, which sufficiently removed the day-to-day bias from the current set.

The individual spectra from the pure milk powders measured in triplicate showed a relatively large number of outliers as determined using appropriate multivariate outlier detection (Mahalanobis distance [18] & PCA-sample residuals). In addition, the distance of triplicate measurements in the context of the complete milk powder sample variation (PCA) was evaluated, and where one of the triplicates showed an extreme distance (a rather subjective approach, for now) to the other two, it was flagged as an outlier. These two outlier removal mechanisms did not exclude all triplicates of one sample per sensor type, nor multiple of the sensor types per sample, and are thus likely to be rather random effects. The total quantity of outliers that were flagged as outliers was ~10% of the individual spectra. This is considered rather large, and in the future repeatability of analysis should be improved.

Detection of food fraud requires a special chemometric strategy. Essentially, an accurate quantification of a specific adulterant is not required; it is the presence or absence of an adulterant that should be established. This makes this task a classification problem rather than regression that is used for other use cases. Moreover, where many popular classification models can be trained to detect a specific adulterant in a specific product (for example melamine in milk powder, or urea in milk powder), we would like the model to detect a broad range of adulterants, even for compounds that the model was not trained for. Therefore, we would like to use one-class classification (OCC), where we train the model to recognize the normal (spectral) variation for a pure product, and flag samples as (possibly) adulterated when observed spectra are outside this normal variation.

With building one-class classification models, there are several ways to pre-process the data and several classification algorithms. Since it is impossible to know, *a priori*, which combination of pre-processing and algorithm is best to describe the natural variation of pure products well while at the same time is sensitive enough to detect several forms of adulteration, we decided that in this phase of the feasibility study we should combine a multitude of these options. We also explored the option of spectrum splitting, in which only parts of the (suitable pre-processed) spectrum is subjected to classification, to make the model more sensitive to those type of adulterants that only influence very small regions of the spectra. This type, think of bulk chemicals, might not be detected in full-spectrum OCC as the majority of the spectrum will be very similar to the pure product. The following table lists the steps taken (Table 1):

Table 1 – Data processing chains for SMPs

Stage	Options – one of the options per box is used
Spectrum	• Fluorescence • Visible • Near Infrared
Pre-processing	• SNV (Signal Normal Variate) • SNV detrend; SNV followed by a linear de-trending of the signal • First derivative; 11 points • Second derivative; 11 points • Wavelet transform; haar and 8-Least Asymmetric filters; medium level coefficients retained
Spectrum splitting	• None (full spectrum) • 4 Equal parts • 8 Equal parts
OCC algorithm	• SIMCA (Soft Independent Modelling of Class Analogy), 3 or 4 latent variables • k-NN (k-Nearest Neighbors), orthogonal distance to 2nd nearest • PCA residual (Principal Component Analysis), sample residual using 3 principal components • Mahalanobis distance • OCSVM (One-Class Support Vector Machine), radial kernel

For each combination of the stages listed in the table, the dataset was subjected to repeated random cross-validation. In each of 50 iterations, 80% of the pure milk powder samples were randomly selected to build the OCC model. The 20% complementary milk powder samples and all adulterations were predicted – the class distance was calculated - on this model. The final class distances were averaged of the 50 cross-validation iterations.

The current approach in this feasibility study resulted in 1152 class distances for each sample. Although calculation time to do so is certainly not excessive, this number can and will be reduced, retaining the model's capability in the next phase.

4.2 Data fusion strategy

The 1152 class distances obtained, in a way, represent figures the spectra's typicality seen from many different angles. Even just one extreme value, or a few that are on the far end of a normal distribution, means that the sample (or at least the spectra obtained from it) really is different from the known variation of the pure product. This means that is should be investigated further for adulteration or possible other compositional/quality issues.

The 1152 initial class distances are too many to review manually or to make a consumer-friendly information figure. We need a way to summarize the 1152 individual class distances in a way that does right to the fact that one extreme or several borderline results should flag the sample as deviant. In this phase of research, we opted for a one-class SIMCA model [19], which is built based on the class distance scores for the pure milk powder samples. The class distances of all samples are calculated accordingly, and these values have been presented as "results" in D3.2. This process of data fusion will be refined in next steps of this project.

5 Integration of algorithms in the PhasmaFOOD software platform

The PhasmaFOOD software platform is distributed on three system layers: 1. Embedded system hosted on the PhasmaFOOD smart sensing device; 2. Mobile application hosted on end user smartphone/tablet; and 3. Cloud platform hosted on commercial IaaS (Infrastructure as a Service) platform. It is specified and developed to provide the following functionalities:

- To provide a graphical user interface for end users, enabling them to configure measurement process, manage measurement data and receive analysis results in required format;
- To establish communication channels for exchanging measurement data, configuration commands and analysis results;
- To store all collected use case datasets and make them available for analysis and decision-making model training;
- To properly drive sensors integrated into the smart sensing device and collect measurement data related to project use cases;
- To preprocess collected measurement data and prepare them for analysis in line with project use cases' objectives;
- To implement and execute decision-making algorithms based on data analysis and machine learning models in line with project use cases requirements;
- To support system calibration and management including calibration of data analysis and machine learning models.

In this section, we will provide a description of how the data analysis and decision making algorithms are implemented as part of the PhasmaFOOD software platform. More detailed description of the software components can be found in D4.1.

5.1 Requirements for implementation of the data analysis chains and detection algorithms

The process of specification, implementation and validation of data analysis and decision-making procedures are performed in line with requirements defined in D1.2. Table 2 lists all requirements for guiding implementation of the data analysis chains and decision-making models in the PhasmaFOOD system and software platform.

Table 2 - Requirements for data analysis and decision making processes for realization of the PhasmaFOOD use cases

Ref. number (see D1.2)	Description	Implementation in the first system prototype
DATA-CALIB-3	Data analysis algorithms and related protocols deployed on the cloud platform SHOULD provide calibration recommendations based on specific application.	The decision-making procedures which are hosted on the cloud platform, decide on proper data handling and system calibration. Mechanisms for online model calibration will be investigated for the final system prototype. Also, during the validation and experimentation sessions, we will define best practices to be included into recommender system.
DATA-CALIB-4	Data analysis algorithms and related protocols on the PhasmaFOOD device/smartphone/cloud SHOULD provide 'online' system training/calibration capability to improve on the 'offline' or 'user-initiated' calibration process.	
DATA-COMPRESS-1	Embedded µ-controller software (and related hardware specification, i.e. µ-controller, memory etc., specifications) on PhasmaFOOD device SHOULD implement compression tools for 'onboard' pre-processing of sensor data.	Camera image compression techniques are in process of implementation.
DATA-COMPRESS-2	NIR sensor data rates (expected below Mb/s range) SHOULD be pre-processed ('filtered') in conjunction with the sensor device capability to reduce noise margin and/or possibly extract relevant signal components exploiting learned or known prior information.	The embedded software performs preprocessing – normalization with dark and white reference and averaging. In the cloud platform data analysis chains perform feature selection and dimensionality reduction.
DATA-COMPRESS-3	Embedded software components MAY provide in addition feature extraction capability.	The first prototype of the embedded system does not include feature extraction.

DATA-COMPRESS-4	UV-VIS data rates (expected in Kb/s range) SHOULD be pre-processed the same way as NIR sensor data.	See above
DATA-COMPRESS-5	Camera data SHOULD be compressed using standard image compression tools.	Image compression algorithms are selected and under implementation (see D5.3)
SW-ARCH-5	The PhasmaFOOD system SHOULD offer hybrid processing and data analysis capabilities between the PhasmaFOOD device (accelerators, microcontrollers), a mobile application on a smartphone/tablet and the PhasmaFOOD cloud system.	The first prototype implements data preprocessing on the embedded system and data analysis with classification and regression models in the cloud platform.
SW-ARCH-8	The PhasmaFOOD SW architecture MUST provide reactive decision making based on sensory data analysis.	The first prototype implements reactive decision making as a rule engine running on the cloud platform. It employs proper data analysis pipeline based on measurement configuration provided by the end user.
SW-EMBED-8	The embedded software SHOULD perform data preprocessing operations.	Embedded system in the first prototype performs data normalization and applies dark and white references.
SW-APP-4	The PhasmaFOOD mobile applications SHOULD be able to perform additional data compression, filtering and analysis.	In the first prototype, the mobile application acts as a communication interface between the embedded system and the cloud platform and towards the end user. Data analysis and decision

		making will be explored for the final prototype.
SW-APP-5	The PhasmaFOOD mobile application MAY be able to host a trained machine learning model for certain Use Cases of food analysis.	Will be explored for the final prototype.
SW-CLOUD-1	The PhasmaFOOD cloud platform MUST host trained decision making machine learning algorithms.	See section 5.2 of this document.
SW-CLOUD-2	The PhasmaFOOD cloud platform MUST host calibration algorithms for the PhasmaFOOD system.	In the first prototype, the calibration operations are performed manually through the command line interface on the cloud platform.
SW-CLOUD-3	The PhasmaFOOD cloud platform MUST perform offline training for machine learning models based on collected data sets.	Machine learning models include regression and classification. Models are trained for each available data set. The web dashboard will include an interface which will enable expert users to train and validate different classification and regression algorithms on the same dataset.
SW-DASHB-1	The PhasmaFOOD cloud platform SHOULD provide web dashboards for system monitoring, calibration, configuration and protocol.	The first version of the web dashboard allows users to navigate and query the cloud database. Further development will include machine learning “playground” and model calibration procedures and interface.
PAR_2	The PhasmaFOOD mobile applications and web dashboards MUST provide an interface for parameter settings specific for each Use Case.	The mobile application provides interface for specifying measurement configuration parameters

		including use case, sample, sample state and configuration of individual parameters.
PAR_3	The mobile application MUST provide GUI for end users allowing them to specify conditions in which the measurements are conducted.	See above.

5.2 Implementation of data analysis chains and integration with a decision-making framework

The implementation of the data analysis procedures, as well as decision-making procedures as part of the PhasmaFOOD software platform, are a vital part of the process for on-site application of the three use cases. The procedures for the implementation of this procedure are extensively addressed in D4.1 and consider:

- Data analysis and machine learning pipelines;
- Accessing the training database;
- Specification and selection of DA/ML (Data Analysis and Machine Learning) models;
- Model training and storing;
- Decision-making algorithms.

5.3 Calibrating smart algorithms

Calibration of decision-making procedures and DA/ML pipelines is performed manually by software developers and validators in the first system prototype implementation. For the final prototype, we will work on implementing automatic procedures for calibration.

All datasets are divided into training and test portions. This was achieved by employing a random splitting approach, based on a configurable percentage value that splits the available data into test and training sets when they are loaded in the cloud database. While for prototype tests a 70/30 ratio was used resulting in 70% training data and 30% test data, choosing other split ratios can easily be achieved by the implemented tools. The training portion was used in the PhasmaFOOD cloud platform to train the machine learning models for classification and regression, which are part of the decision-making mechanisms for project use cases. The test/validation portions of these datasets were used to emulate real measurements, coming from the integrated PhasmaFOOD smart sensing device. This is the first step in validating trained decision-making models.

For extracting the test data in a suitable format and to ensure in the process that training data will never be used in subsequent tests, a set of scripts was developed that serves the exact purpose of extracting the test data from the cloud database (30% of the data sets) and generating files identical to the ones that would have been produced by the actual sensor hardware, for all sensor types. In addition to the test data, the extraction scripts also provide the classification information of the samples in a suitable format together with relative information (contamination levels, aflatoxins, and microbiological data), providing adequate information to simulate a measurement and verify that the correct classification response was subsequently received. It is also worth mentioning that specifically for the cases, were averaging and/or preprocessing is to be performed on the embedded device (e.g. average across measurements, Dark subtraction and White division), the averaged and pre-processed values are also calculated and available in the cloud database. This provides a mechanism to further ensure, if necessary, the fact that pre-processed and averaged values originating from the embedded device are matching to the ones that were calculated by a different implementation on a different platform.

The performance of decision-making algorithms is tested before deployment (integration in the cloud platform processes) using the dedicated test data for each dataset. For classification tasks, the function *sklearn.metrics.classification_report()* is then applied to the ground truth and the predicted labels to obtain precision, recall and f-score of the classification. A confusion matrix is calculated with the function *sklearn.metrics.confusion_matrix()*, showing the number of correctly classified samples on the main diagonal and the number of samples from class *i*, wrongly classified as belonging to class *j* in the entry *(i,j)*. For regression tasks, the mean squared error, the mean and median absolute errors and the explained variance are calculated using the respective functions *sklearn.metrics.mean_squared_error(), sklearn.metrics.mean_absolute_error(), sklearn.metrics.median_absolute_error()* and *sklearn.metrics.explained_variance()*. A standardized test suite is currently being developed and will be part of the web dashboard.

6 Conclusions and outlook

In this work, the activities on developments on detection algorithms until June 2018 (M18) are reported. Each use case has its own unique chemometric strategy and data-fusion approaches in order to utilize the three unidimensional data-sets produced by the PhasmaFOOD optical sensors. For each use case, promising initial results have been achieved by integrating data streams for detecting of aflatoxins, detection of food spoilage and adulteration. Furthermore, in conjunction with WP4, an integration strategy is proposed for integration of the smart algorithms in the PhasmaFOOD cloud.

In this phase of the development of the detection algorithms, data was used which was recorded using the individual PhasmaFOOD sensors, *i.e.* prior to the deliverance of the PhasmaFOOD integrated prototype. During the final course of WP3, the PhasmaFOOD prototype will be used for the further building of spectral databases and validation of the detection algorithms and data-fusion strategies as described in this document. Furthermore, the detection algorithms will need to be validated and calibrated, a process which will be continued in WP6 as WP3 will finish in March 2019.

7 References

1. Magna, G., et al., *Unsupervised On-Line Selection of Training Features for a robust classification with drifting and faulty gas sensors.* Sensors and Actuators B: Chemical, 2018. **258**: p. 1242-1251.
2. Esteban, J., et al., *A Review of data fusion models and architectures: towards engineering guidelines.* Neural Computing & Applications, 2005. **14**(4): p. 273-281.
3. Khaleghi, B., et al., *Multisensor data fusion: A review of the state-of-the-art.* Information Fusion, 2013. **14**(1): p. 28-44.
4. Brown, D.S., *Chemometrics: A textbook. D. L. Massart. B. G. M. Vandeginste, S. N. Deming, Y. Michotte, and L. Kaufman, Elsevier, Amsterdam, 1988. ISBN 0-444-42660-4. Price Dfl 175.00.* Journal of Chemometrics, 1988. **2**(4): p. 298-299.
5. Martens, H. and T. Naes, *Multivariate Calibration*, in *Encyclopedia of Statistical Sciences*. 1989.
6. Duda, R.O., P.E. Hart, and D.G. Stork, *Pattern Classification (2nd Edition)*. 2000: Wiley-Interscience.
7. Barnes, R.J., M.S. Dhanoa, and S.J. Lister, *Standard Normal Variate Transformation and De-Trending of Near-Infrared Diffuse Reflectance Spectra.* Applied Spectroscopy, 1989. **43**(5): p. 772-777.
8. Larive, C.K. and J.V. Sweedler, *Celebrating the 75th Anniversary of the ACS Division of Analytical Chemistry: A Special Collection of the Most Highly Cited Analytical Chemistry Papers Published between 1938 and 2012.* Analytical Chemistry, 2013. **85**(9): p. 4201-4202.
9. Wold, S., M. Sjöström, and L. Eriksson, *PLS-regression: a basic tool of chemometrics.* Chemometrics and Intelligent Laboratory Systems, 2001. **58**(2): p. 109-130.
10. Martens, H., et al., *Analysis of designed experiments by stabilised PLS Regression and jack-knifing.* Chemometrics and Intelligent Laboratory Systems, 2001. **58**(2): p. 151-170.
11. Breiman, L., *Random Forests.* Machine Learning, 2001. **45**(1): p. 5-32.
12. Guo, Q., W. Wu, and D.L. Massart, *The robust normal variate transform for pattern recognition with near-infrared data.* Analytica Chimica Acta, 1999. **382**(1): p. 87-103.
13. Duffy, N. and D. Helmbold, *Boosting Methods for Regression.* Machine Learning, 2002. **47**(2): p. 153-200.
14. Tin Kam, H., *The random subspace method for constructing decision forests.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 1998. **20**(8): p. 832-844.
15. Friedman, J.H., *Greedy Function Approximation: A Gradient Boosting Machine.* The Annals of Statistics, 2001. **29**(5): p. 1189-1232.

16. Panagou, E.Z., et al., *Potential of multispectral imaging technology for rapid and non-destructive determination of the microbiological quality of beef filets during aerobic storage.* International Journal of Food Microbiology, 2014. **174**: p. 1-11.
17. Papadopoulou, O., et al., *Contribution of Fourier transform infrared (FTIR) spectroscopy data on the quantitative determination of minced pork meat spoilage.* Food Research International, 2011. **44**(10): p. 3264-3271.
18. De Maesschalck, R., D. Jouan-Rimbaud, and D.L. Massart, *The Mahalanobis distance.* Chemometrics and Intelligent Laboratory Systems, 2000. **50**(1): p. 1-18.
19. Brereton, R.G., *One-class classifiers.* Journal of Chemometrics, 2011. **25**(5): p. 225-246.