Comparison of different data and information fusion methods to improve the performance of machine learning models

Pál P. Hanzelik\*a,b, Alex Kummerb, Márton Móczc, Szilveszter Gergelyc, Dorián L. Galatad, János Abonyib

aGroup Enterprise Data, MOL ITD GBS Hungary Ltd., Dombóvári út 28., H-1117 Budapest, Hungary

bHUN-REN Complex Systems Monitoring Research Group, University of Pannonia, Egyetem u. 10., H-8200 Veszprém, Hungary

c Dept. of Applied Biotechnology and Food Science, Faculty of Chemical Technology and Biotechnology, Budapest University of Technology and Economics, Műegyetem rkp. 3., H-1111 Budapest, Hungary

d Dept. of Organic Chemistry and Technology, Faculty of Chemical Technology and Biotechnology, Budapest University of Technology and Economics, Műegyetem rkp. 3., H-1111 Budapest, Hungary

\*[pphanzelik@molitdgbs.com](mailto:pphanzelik@molitdgbs.com)

Abstract

The combined handling of information from fast and non-destructive spectroscopic measurement techniques enables more accurate and robust models. Machine learning algorithms that estimate qualities are sometimes not accurate enough for a given quality, so to improve the models' accuracy, we need to provide additional information from different analytical measurements. This research compares five data and information fusion techniques tested on spectroscopic results from various oil industry samples. We aim to show examples of which fusion technique should be used and how to build models in this way. We used the boosting technique to compare which approach provides the most information from the measurements. We used mid-infrared and Raman spectral data from the same samples in this study and applied low, medium, high, and complex data fusion techniques. Our motivation is to combine and compare the treatment of different measurement techniques that can provide additional information. The difference between the five methods is the level at which the fusion takes place. The variables come from spectral data at a low level, but at a medium level, we used individually created model results. Furthermore, at the complex level, we used the data of the models built together with the ensemble learning technique for the spectral data.

Considering our achievements in the oil industry, data fusion techniques can significantly improve the accuracy of machine learning models. In our case study, the best results were obtained by the fifth-level data fusion technique, where half of the developed model can predict the hydrocarbon/imide ratio of the additive with an error as if data fusion had not been applied.

**Keywords**: model development, data fusion, ensemble learning, spectroscopy, quality control

* 1. Introduction

By storing and using the data from different measurements in a shared database, we can provide data faster and more efficiently. We can obtain more accurate models with machine learning (ML) algorithms built on fused data. Based on increasingly essential quality criteria and environmental protection regulations, data fusion techniques built on laboratory data represent a vast potential for digitizing laboratory activities. The quality of a material can be determined from the results of several laboratory measurements. However, to reduce the turnaround time of the laboratories, fast measurements with much information are preferred. According to the literature review, the data fusion techniques began to spread in the second half of 2017, after which they gained more and more emphasis in the work of quality assurance laboratories. Most scientific literature on data fusion techniques was published in the food and pharmaceutical industries. The growing number of multi-dimensional analytical instruments and chemical measuring devices that produce data sets of different dimensions provide a considerable amount of analytical information about complex samples and, at the same time, pose a challenge during data evaluation. The use of ML tools helps the efficient use of these data volumes. Furthermore, the development of chemometrics revolutionized the steps in the interpretation of analytical processes and contributed to the solution of more complicated analytical problems.

Data fusion techniques are necessary to examine the complexity of an analysis problem from several different perspectives. The goal is to create the most comprehensive and complex picture possible of the reviewed materials, and it is also essential for us to obtain as much information as possible from the point of view of qualification quickly and non-destructively using a small number of samples. Many questions and challenges in food chemistry or the pharmaceutical industry can only be solved by combining the results of different instruments with a broader characterization of the given samples. Data fusion enables the simultaneous extraction of meaningful and valuable information from various analytical sources. It can be seen from the literature that in recent years, data fusion has produced better results for the investigated systems than the evaluation of individual data sets [1]. Spectroscopic measurements result in multiple data points whose historical storage and use in modelling provide an opportunity to improve the quality assurance of production processes with an Industry 4.0 approach. Data fusion techniques were applied to modern process analytical technology (PAT) data for non-destructive prediction of drug tablet dissolution. They could directly model the dissolution of the tablets from the near-infrared (NIR) and Raman spectra of the tablets. Data fusion techniques have greatly improved the accuracy of the models. During the research, partial least squares (PLS) models were built on the data of the NIR and Raman devices separately or in combination, and artificial neural network models were constructed on their output to predict better dissolution by Nagy et al. [2]. In the food industry, for example, research indicating the caprolactam content of sauce-based foods is also an excellent example of applying data fusion techniques. The measurements were performed using near-infrared (NIR) and mid-infrared (MIR) spectroscopy, and models were built using the data fusion technique (PLS) and support vector machine (SVM) based on the measurement results. The results of each method were combined into different weighted coefficients according to their contribution by the work of Zhu et al. The high-level fusion model provided the best performance, followed by the MIR model, then the low, and medium-level fusion models, and finally, the NIR model proved to be the weakest [3]. In our research, we tested different data fusion approaches on the infrared and Raman spectroscopic measurement spectra on oil industry samples to build models with sufficient accuracy for quality parameters that are difficult to predict. In the case study, we present an additive's hydrocarbon/imide ratio, where the model built on the Raman and MIR models needed sufficient accuracy in terms of proper classification.

* 1. Methodology

We performed the IR and Raman spectroscopy measurements of the same samples. Then, we built ML models to predict the sample quality using different data fusion techniques to predict the hydrocarbon/imide ratio of the additive. Our goal is to demonstrate that using and properly producing quality data from other analytical measurements can result in more accurate ML models.

Figure 1 presents the five investigated model development techniques with no or different data fusion techniques. In mode *A*, we built a model separately for the IR and Raman data. In this case, we did not apply data fusion. We did this for reference, to which we can compare the results of different methodologies, and the other four modes are distinguished depending on the implementation level of data fusion techniques. In the low-level data fusion (*B*) route, the raw datasets were fused and analyzed as a single dataset. In this case, feature selection was applied to the entire data set, and then a regression ML model was built. The middle-level data fusion path (*C*) fuses important actual or latent variables extracted from feature selection on each data set. We combined the essential properties of infra and Raman and built a model. The high-level data fusion path (*D*) builds a separate model for all available data sets, combines the model prediction results, and then merges them to create a standard model using the ensemble technique. The complex-level data fusion path (*E*) is an ensemble model built on the latent variables of the models based on the pre-processed and fused data set and the model prediction results.

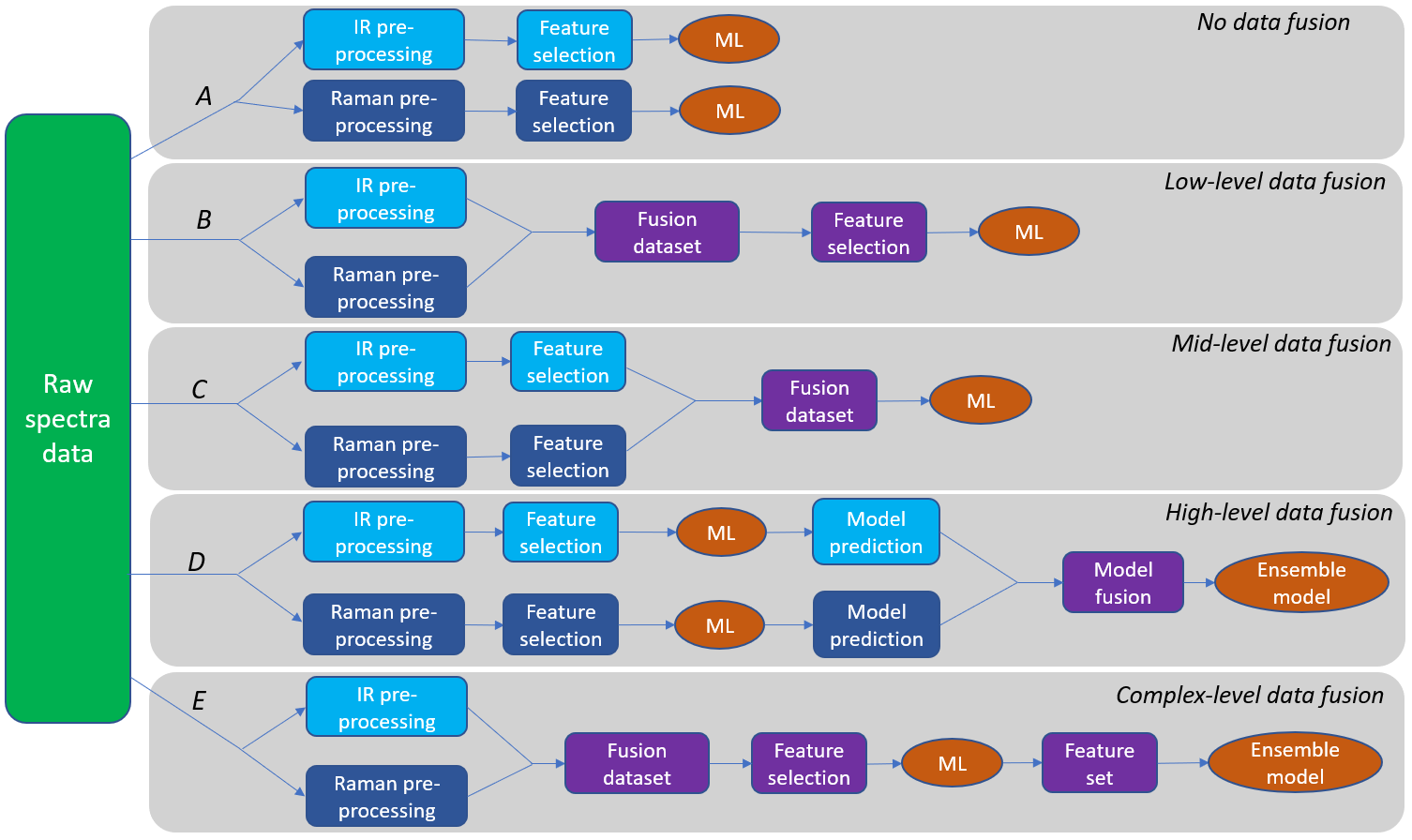


Figure 1: Five different techniques of data fusion.

The essence of the ensemble model used in ways *D* and *E* is that it performs the prediction by combining the results of the previously built ML model or models and considering its prediction. Ensemble learning has become one of the practical and popular techniques for increasing the accuracy of ML models in data science. By applying the methodology, we aim to emphasize the strength of individual ML models and achieve more accurate and robust results.

The feature selection in Figure 1 modes *A*, *B*, *D* and *E* refers to the Genetic Algorithm (GA). This technique automatically selects the variables that result in models with lower root-mean-square error (RMSE) values. GA runs were performed with a window width of 20 and a population size of 64.5 replicate runs, as each time, the initially included variables are selected randomly. Therefore, the results can vary each time.

In the case of mode *C*, feature selection was the Principal Component Analysis. The value of the first ten principal components was used to represent the spectra. These principal components describe more than 90% of the total variance for both Raman and infrared.

The ML model in Figure 1 refers to PLS models combined with GA. The modes *D* and *E* in the ensemble learning technique have different versions depending on the method used for learning, but in this study, we dealt in detail with the boosting technique. We chose the eXtream Gradient Boost (XGBoost) technique in this research because the algorithm can dynamically determine the depth of the decision trees used as weak learners, adding penalty parameters to the high-depth prevention trees, which can prevent the model from overfitting or improving its performance [4].

* 1. Case studies

Our goal is to predict the quality of additives accurately. For this, we need fast, easy-to-use and accurate forecasting. We preferred fast, non-destructive techniques based on spectral data, but if only IR or Raman techniques were used, the accuracy of the models needed to be improved. Therefore, we need to use data fusion to improve quantitative forecasting. In our case, this is the hydrocarbon/imide ratio of the selected additive. During the production of additives, the classification is based on chemical and physical parameters. Quality parameters are difficult to predict, so we have to use data fusion techniques. In the future, our goal is to integrate inline Industrial Internet of Things (IIoT) sensors into our processes in addition to laboratory measurements, which can be used to predict the quality of raw materials, intermediate products, and final products online in real-time, as well as the different effects of quality on production [5]. Our chosen analytical measurements are spectroscopic measurements, including IR and Raman spectroscopy. Both measurement techniques have industrial sensors available on the market.

We used IR and Raman spectroscopy in the research because they are complementary analytical tools. During Raman inelastic light scattering, the monochromatic excitation laser beam hits the sample material, and the scattered light provides information about molecular vibrations and chemical structure. Raman detects vibrations along the covalent bonds, the advantage of which is that the possible water content of the sample does not interfere with the measurement, in contrast to IR spectroscopy. IR is based on the molecular absorption of irradiated IR light caused by vibrational and rotational transitions in covalent bonds. The IR range is the spectral section between 12,800 and 10 cm-1, within which we distinguish three different regions, and the mid-infrared range (MIR: 4000-650 cm-1) was used in this study. At the same time, infrared detects the vibration of polar covalent bonds, while Raman detects the vibration of non-polar covalent bonds. The advantage of infrared spectroscopy is that it is not disturbed by fluorescence compared to the Raman technique [6]. Our target variable is the hydrocarbon/imide ratio, an essential parameter when qualifying the additive sample. Our goal was to build an accurate model for this parameter when building the model, so we compared the five modes based on these.

* 1. Results

The number of additive samples was 99, whose IR and Raman spectra were measured. The dependent variable X of the spectrum of the samples was preprocessed with Autoscale and Mean center settings in the PLS toolbox of MATLAB 2020a.

The results were summarized based on the methodology presented in Section 2; the models were compared based on RMSE and Pearson R2 key performance indicators. The models were checked using both calibrations, where all samples were used for model development and the Venetian blinds cross-validation techniques. In the case of the Venetian blinds cross-validation, each test set is determined by selecting every s-th object in the data set, starting at objects numbered 1 through s, which in our case was s=10 (Figure 2, 3).

Figure 2: The square of Pearson correlation coefficient of the five modes: *A*, *B*, *C*, *D* and *E*.

Figure 3: RMSE of the five modes: *A*, *B*, *C*, *D* and *E.*

The results show that modes *B*, *D* and *E* gave better results than mode *A*. Among the five methods, *E* gives the best results. Venetian blinds Cross-validation reduces the model error (RMSE) by more than half, doubling the square of Pearson correlation (R2). In mode *E*, we processed the input spectra separately for IR and separately for Raman, then pooled the entire spectral data set and ran PLS regression. After the PLS model, we built an ensemble model, which was an XGBoost regression model. The latent variables of the first model served as input, which were selected by the generative algorithm. The weakest results were given by mode *C*, with medium-level data fusion. The results were even lower than in mode *A*, where we did not use data fusion. Mid-level data fusion (mode *C*) performs feature selection on infrared and Raman data sets. After selecting the function, we compiled the reduced data set produced by the PCA analysis of the IR and the Raman data set separately. We then built a PLS regression model for the 10-10 main components. The results of modes *B*, *C*, *D*, and *E* prove that selecting variables is essential when using data fusion techniques..

* 1. Conclusion

The results show that the models based on the hydrocarbon/imide ratio of the investigated substance with the low, high and complex data fusion techniques gave better results than those built on only the MIR or the Raman data. The results also show that the intermediate technique result was worse than the original one when we did not use the data fusion technique. When applying the methodology, selecting variables is essential in all modes *B*, *C*, *D* and *E*. This selection can be done manually or automatically, and in our use case, we used the genetic algorithm to select the features. The mode *E* gives the best result as a complex data fusion technique. Using complex-level data fusion (*E*), Pearson R2 is 0.996, RMSE is 0.154 from the calibration set, Pearson R2 is 0.850, RMSE is 0.975 from Venetian blinds cross-validation. The mid-level data fusion tested on *C* did not give better results than when we built a separate model for the two data sets. The weaker result of mode *C* can be explained by the fact that when applying the data fusion technique, the selection of features must be handled together, not separately.

In summary, the methodology presented in this study is suitable for building more accurate ML models based on Raman and IR datasets. When developing models, it is essential to be careful and to filter out noisy spectral ranges. In connection with the development of the model, we can extract and combine information from other sources to improve the performance of the models. We can further enhance the performance of the ML models by directly using sensor data from industrial plants (e.g. temperature, pressure) to estimate product quality.

**Acknowledgements**: The research was funded by the National Research, Development and Innovation Fund in the frame of the 2019-1.3.1-KK-2019-00004 project.

References

[1] A. Biancolillo, F. Marini, C. Ruckebusch, R. Vitale, 2020, Chemometric strategies for spectroscopy-based food authentication, Applied Sciences, 10, 18, 6544.

[2] B. Nagy, D. Petra, D. Galata, B. Démuth, E. Borbás, G. Marosi, A. Farkas, 2019, Application of artificial neural networks for Process Analytical Technology-based dissolution testing, International Journal of Pharmaceutics, 567, 118464.

[3] J. Zhu, X. Fan, L. Han, C. Zhang, J. Wang, L. Pan, M. Zhang, 2021, Quantitative analysis of caprolactam in sauce-based food using infrared spectroscopy combined with data fusion strategies, Journal of Food Composition and Analysis, 104, 104130.

[4] A. Kumar, J. Mayank, 2020, Ensemble learning for AI developers, BApress: Berkeley, CA, USA, 2-56.

[5] P. P. Hanzelik, A. Kummer, J. Abonyi, 2022, Edge-Computing and Machine-Learning-Based Framework for Software Sensor Development, Sensors, 22, 11, 4268.

[6] Monograph. NIR Spectroscopy, 2014, A guide to near-infrared spectroscopic analysis of industrial manufacturing processes, Metrohm AG, CH-9101 Herisau, Switzerland.