Trustworthiness of laser-induced breakdown spectroscopy predictions via simulation-based synthetic data augmentation and multitask learning

Riccardo Finotello, Daniel L'Hermite, Céline Quéré, Benjamin Rouge, Mohamed Tamaazousti, Jean-Baptiste Sirven

Trustworthiness of Laser-Induced Breakdown Spectroscopy Predictions via Simulation-based Synthetic Data Augmentation and Multitask Learning

Riccardo Finotello[1,*], Daniel L'Hermite[2], Celine Quéré[2], Benjamin Rouge[2], Mohamed Tamaazousti[1], and Jean-Baptiste Sirven[2]

[1]Université Paris-Saclay, CEA, LIST, Palaiseau, F-91120, France
[2]Université Paris-Saclay, CEA, Service de Physico-Chimie (SPC), Gif sur Yvette, F-91191, France
(*) riccardo.finotello@cea.fr

***Abstract*—Laser-induced breakdown spectroscopy is a versatile technique that can be used to quickly measure the concentration of elements in ambient air. We tackle the issues of performance and trustworthiness of the statistical model used for predictions. We propose a method for improving the performance and trustworthiness of statistical models for LIBS. Our method uses deep convolutional multitask learning architectures to predict the concentration of the analyte and additional information as auxiliary outputs. We also introduce a simulation-based data augmentation process to synthesize more training samples. The secondary predictions from the model are used to characterize, quantify and validate its trustworthiness, taking advantage of the mutual dependencies of the weights of the neural networks. As a consequence, these output can be used to successfully detect anomalies, such as changes in the experimental conditions, and out-of-distribution samples. Results on different types of materials show that the proposed method improves the robustness and trueness of the predictions.**

***Keywords* —LIBS; trustworthy AI; simulation; multitask; neural networks.**

I. Introduction

DEEP Learning (DL) has been adopted with success in many areas of science, and analytical chemistry is no exception. Laser-Induced Breakdown Spectroscopy (LIBS) offers plenty of applications for MVA techniques, especially for hyperspectral imaging applications [1], [2], but the development of DL techniques is quite recent. Basic shallow Neural Networks (NNs) were implemented as soon as the '90s for identification of polymers [3]. The introduction of modern concepts and architectures of DL is however quite recent [4].

In spectroscopic quantitative analysis, we build models relating experimental spectra to the concentration of the species of interest, using calibration samples. Quantitative models are then valid for a given sample matrix and for given experimental conditions. However, for direct analytical techniques like LIBS, unknown samples can have different matrices, and experimental conditions can change. The predicted concentration can thus be biased to a certain extent. Yet, models are usually designed to deliver a concentration but not to estimate to which extent an unknown sample is well represented by calibration ones. In other words, we do not know how reliable the prediction is. Hence, we need models that are able to estimate, if possible quantitatively, the trustworthiness of a prediction. Thanks to their ability to perform different tasks in parallel, MultiTask (MT) NNs are a type of model that could address this issue, by providing, at the same time, the main expected prediction and a characterization of its properties. In turn, these auxiliary data can be used to validate and quantify the model trustworthiness. However, deep NNs can easily overfit the training data, since they require a very large number of spectra to be properly trained: experimentally, time and cost constraints generally do not enable to acquire enough spectra. A possibility to overcome this limitation is to generate calibration data. This can be done by phenomenological modelling of training spectra.

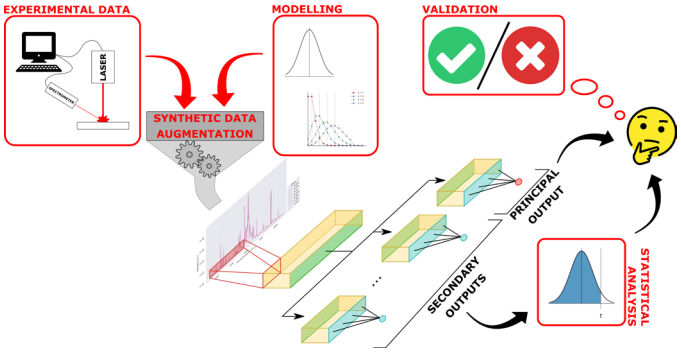

Fig. 1. A robust multitask model is trained on synthetic data to predict a principal variable (e.g. the concentration of the analyte) and several secondary spectral quantities.

We introduce a new approach of quantitative analysis by LIBS and DL, enabling both to generate enough spectra to train a model based on MT Convolutional Neural Networks (CNNs), and to evaluate the confidence in the model predictions. Namely, we propose:

1. a simulation-based synthetic data creation, that is the synthesis of an arbitrary number of new spectra from experimental data to train complex DL architectures;

2. robust deep MT CNNs, capable of processing entire LIBS spectra and increasing the robustness (homoscedasticity) of the model;
3. a measure of the trustworthiness of the predictions of the model through the statistical analysis of the outputs of the MT architecture.

The procedure, presented in Fig. 1, aims at providing a way to increase the robustness of the calibration models, and to benchmark the confidence of the Artificial Intelligence (AI) architecture through its own predictions.

II. Related Works

Univariate and multivariate (MVA) calibration techniques are explored in the LIBS literature [5]. The first are widely adopted for their ability to provide interpretable results quickly and with good precision [6]. In this scenario, calibration standards are used to build a map between the concentration of a given analyte and the information contained in a single measurable variable, such as the integral intensity of an emission line. The model is then inverted during inference to use the measured intensity as predictor of the concentration of the analyte. On the other hand, MVA methods were introduced for their ability to take advantage of more information contained in the input spectra, rather than focusing on a single variable. Techniques based on principal components and multilinear regression have been widely adopted in LIBS [7], [8]. As a MVA technique, DL has also been explored, comparing various types of architectures [4].

A. Data Augmentation

Though DL shows potential for various analyses, time of data gathering and experimental conditions often prevent building large LIBS datasets. Feature engineering and feature selection, for instance through a principal components analysis or a priori expertise, have been employed to reduce the size of the input data in order to be used in smaller or more adapted machine learning (ML) models [9]. At the same time, the production of purely synthetic data, based on local thermodynamical equilibrium, has been explored in LIBS applications [10]. The idea of enriching existing data by means of different representations of the inputs, such as time resolved spectra, was experimented with success [11]. More recently, standard data augmentation techniques were used for the classification of LIBS mapping experiments [4]. Augmentation through synthetic data via the simple addition of random noise to the experimental data is also briefly discussed in the literature [12].

B. Multitask Learning

Though explored at length in ML, MT learning [13] has seen major developments and a wide range of applications recently. Some examples of multi-output algorithms for LIBS analyses were recently explored, based on NNs [11] and on Partial Least Squares (PLS) with two outputs, i.e. PLS2 [14]. In these cases, the response variables are usually the concentrations of multiple elements.

Interestingly, multi-output NN architectures found successful applications for the computation of plasma parameters from LIBS data [15]. MT learning was also introduced for the simultaneous predictions of concentrations of analytes and lithology classes: different NNs were used to process the data with different loss functions, using a latent representation of the input, computed by a common backbone architecture [16].

C. Trustworthy AI

Standard techniques in statistics, such as confidence intervals for mean values and predictions, are usually preferred [17], though they rely on strong assumptions on the type of data analyzed, such as the independence of the residuals from the independent variables. On the other hand, it has been seen heuristically that traditional models do not systematically generalise to unknown data. Some Explainable-AI (xAI) methods have been introduced to analyze the progression of the feature maps in the hidden layers [18], or study the importance of the spectral variables leading to the prediction [19]. This represents indeed a step towards the comprehension of the mechanisms behind NNs. Nevertheless, it does not deal with the confidence of the predictions, or the automatic detection of changes in the distribution of the samples.

III. Methodology

In what follows, we detail the contributions to the end-to-end pipeline of the analysis, from the synthetic data augmentation, to the validation of the model predictions through the statistical analysis.

A. Synthetic Data Augmentation

Training large NNs by optimizing the bias-variance tradeoff may lead to phenomena such as poor stability and bad generalization in inference on unknown samples. Many training samples are usually required to train more complex architectures to overcome the issues. Inspired by usual DL practices, we thus introduce a data augmentation technique for LIBS.

We train the MT model on entire LIBS spectra, without a priori data selection. Given the scarcity of training data, the simple addition of random noise to the experimental spectra may result in a training distribution no longer representative of the use case. Thus, we first proceed to model the distribution of the original spectra, and then to synthetically produce an arbitrary number of spectra. Noise can then be added to each individual channel, provided that its global average effect is negligible, in order to make the synthetic distribution more realistic. The procedure ensures to enlarge consistently the feature space spanned by the synthetic spectra. Specifically, we consider the set of n spectra with p wavelength channels $\{x^{(i)} \in \mathbb{R}^p\}_{i\in[1,n]}$ and the corresponding average spectrum $\bar{\boldsymbol{x}} = (\bar{x}_r)$. For each wavelength channel $r = 1, 2, \dots, p$, we fix the expected value of a random variable $y_r \in \mathbb{R}$ such that $\mathbb{E}[y_r] = \bar{x}_r$. An arbitrary number $m \gg n$ of full spectra can be constructed by similarly proceeding for all wavelength channels. A degree of noise can then be added to each channel, using a multiplicative Gaussian factor $z \sim \mathcal{N}(1, \beta)$, where β is a noise parameter of the synthetic spectra. We can use the

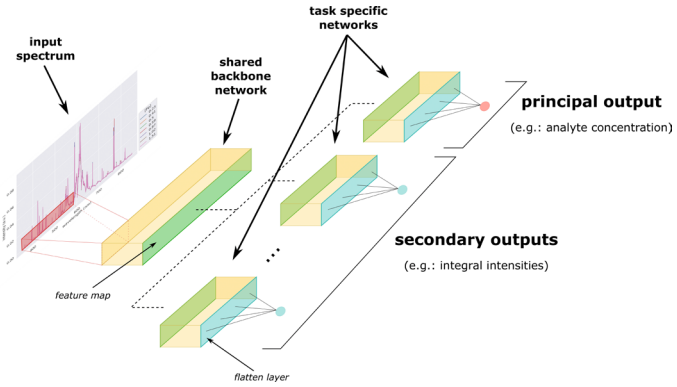

Fig. 2. The MT model is a hard parameter sharing structure, that is a common backbone network and several heads, connected to the shared feature map.

spectra in the training sets to choose appropriate values of β. We observed that, for specific wavelength channels of interest, in the absence of the parameter β, the statistical model representing the spectra is over-confident (i.e. it does not reproduce the full distribution) at low intensity and under-confident at higher intensities (i.e. it creates out-of-distribution intensities), while the opposite occurs for higher values of the noise parameter. This suggests that there is an in-between optimum of the parameter, for which the synthetic distribution covers correctly the variance of the training distribution at a local level. The value of β can be chosen deterministically, for instance, by maximizing the coefficient of determination (R^2) between the ground truths and the synthetic quantiles at a given wavelength of interest.

B. Multitask Convolutional Neural Networks

By definition, MT NNs are a broad class of algorithms, which provide multiple predictions at the same time, using a shared structure of weights, trained simultaneously. This property gives the networks great versatility, as it is capable of using information on one task to improve its generalization. This strategy acts as a regularization and reduces the overfitting of the training data, as the model supposedly learns new representations, which should generalize well on all tasks.

In this analysis, we use a hard parameter sharing implementation of MT learning, with a common set of bottom layers (see the general schematics in Fig. 2). The innermost backbone processes the input spectra and produces a new latent vector representation. The task-specific heads separately use the latent representation as inputs to compute scalar regression outputs (principal and secondary). We use 1D convolutions in the spectral dimension of the data cube as main operation. In what follows, we choose to use the concentration of the analyte as the principal prediction of the network, and the integral intensities of the associated emission lines or bands as secondary outputs of the network. In turn, this helps to stabilize the model and increase its robustness (homoscedastic behavior) during inference. It also provides a set of secondary results, which can be used to validate and quantify the performance of the model and to detect anomalies or out-of-distribution samples.

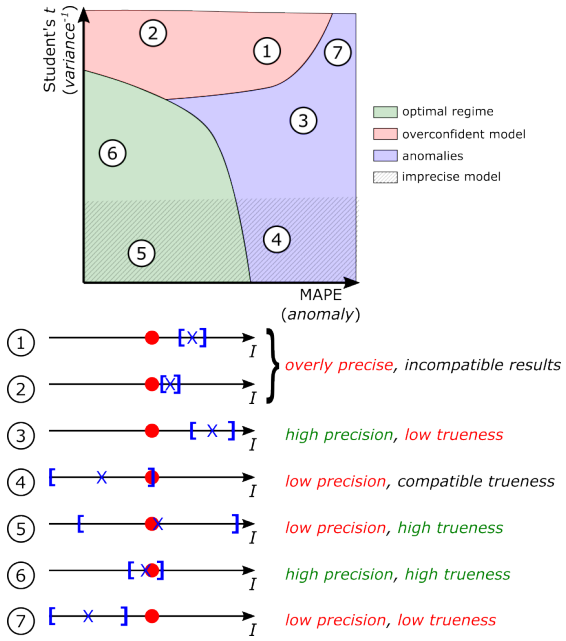

Fig. 3. Outcomes of the MAPE and t variable summarized at the top of the figure. Interpretations are presented schematically at the bottom.

C. Trustworthiness via Validation of the Predictions

In general, the validation of the predictions on new data and the detection of anomalies remain complicated issues in statistics. Given the mutual dependence of the multiple outputs of the MT architecture, we rely on the set of secondary outputs, experimentally measurable on unknown samples. In fact, the integral intensities of emission lines or bands can be extracted from experimental spectra at any given time, even though the concentration of the analyte remains unknown.

When dealing with new data, we compute the predictions of the model on a single spectrum basis, independently. We then average the results per sample, to smooth the influence of defects on the surface. The Mean Absolute Percentage Error (MAPE) gives a measure of the deviation of the data and the performance of the model. For the i-th secondary output of the network and a sample s, let $\boldsymbol{Y}_i^{(s)} = \left|\left(\boldsymbol{I}_i^{(s)} - \hat{\boldsymbol{I}}_i^{(s)}\right)/\hat{\boldsymbol{I}}_i^{(s)}\right|$. We compute the MAPE $M_i^{(s)}$ of the predicted intensities $\boldsymbol{I}_i^{(s)}$ and the corresponding ground truths $\hat{\boldsymbol{I}}_i^{(s)}$, with n samples, that is $M_i^{(s)} \coloneqq \mathbb{E}\left[\boldsymbol{Y}_i^{(s)}\right]$. This offers a first estimate of the error made by the model, as it describes the trueness of the model. In order to estimate a soft threshold for the quantity, we use a validation set of samples. We compute a confidence interval around the MAPE $\left[M_i^{(s)} - \hat{t}_{1-\alpha}^{n}\frac{\sigma_i^{(s)}}{\sqrt{n}}, M_i^{(s)} + \hat{t}_{1-\alpha}^{n}\frac{\sigma_i^{(s)}}{\sqrt{n}}\right]$, where $\left(\sigma_i^{(s)}\right)^2 = \mathrm{Var}\left(\boldsymbol{Y}_i^{(s)}\right)$, and $\hat{t}_{1-\alpha}^{\nu}$ is the value of the *Student*'s t variable for ν spectra, at a confidence level $1-\alpha$.

We then use a Student test on the predicted intensity, given its nature of repeated measurement on the sample. Supposing that the ground truth value of a sample s for the i-th average intensity has a sample variance σ_i^2, we can compute the random variable $t_i^{(s)} = \left|\mathbb{E}\left[\boldsymbol{I}_i^{(s)}\right] - \mathbb{E}\left[\hat{\boldsymbol{I}}_i^{(s)}\right]\right|/\sqrt{\sigma_i^2 + \Sigma_i^2}$, where Σ_i^2 is the sample variance of the predictions. By including the

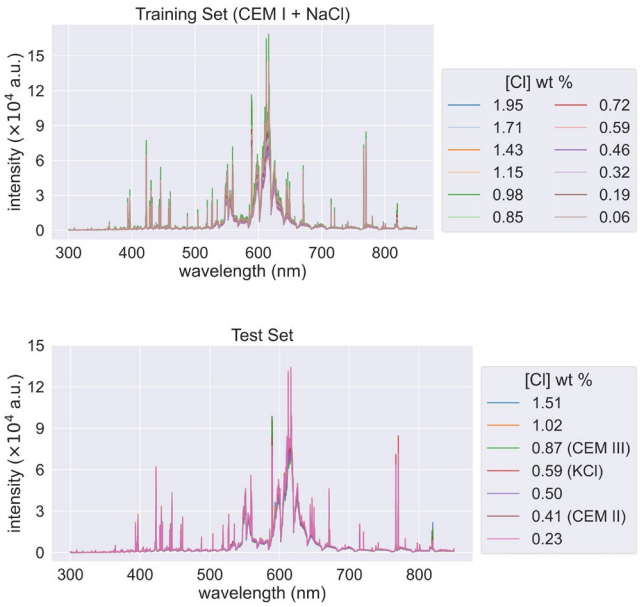

Fig. 4. Average spectra of the cement samples. Samples were fabricated using CEM I, with the addition of NaCl, unless otherwise stated.

dependence on the variance, this test gives a statistical measure of the ability of the model to generalize to unknown data. In order to discriminate possible anomalies, we can use a standard approach by choosing a threshold value $\hat{t}^{\nu}_{1-\frac{\alpha}{2}}$ of a two-tailed t-test at confidence level $1-\alpha$, with ν degrees of freedom, such that the probability $P\left(t_i^{(s)} > \hat{t}^{\nu}_{1-\frac{\alpha}{2}}\right) = \alpha$. This way, we recover a probabilistic interpretation of the result in terms of confidence: models can be compared based on their performance at different values of α on the secondary outputs.

The values of MAPE and of the Student's t variable can be used together to evaluate the trustworthiness of the model and characterize its predictions. We graphically summarize these interpretations in the plane in Fig.3. Though the confidence level of the principal output is not easily computed from the confidence of the secondary outputs, this measure gives an implicit feedback on the main output. Given the dependencies of the MT model parameters, the information determines whether the prediction of the concentration of the analyte is trustworthy.

IV. Experimental Setup

We compare the predictive ability of different algorithms on two types of datasets. We consider 19 cement samples, whose elemental compositions are reported in Fig. 4, and 4 alloy matrices with 4 to 6 samples each, summarized in Fig. 5. The first were built in the framework of an interlaboratory comparison in 2021 [20]. All measurements were carried out in air, at room temperature.

Cement samples were probed using a Nd:YAG laser (*Quantel Brio*) at a wavelength of 1064 nm, 15 mJ pulse energy and 4 ns pulse length. For each sample, 25 spectra were collected, accumulating 40 laser shots (5 pre-ablation shots).

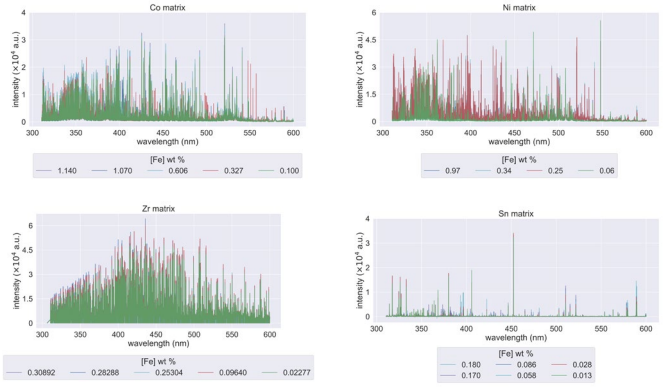

Fig. 5. Different alloy matrices used in the analysis.

We used a *Mechelle* spectrometer with an *Andor iStar* ICCD and a fixed aperture of 50 µm × 50 µm, 10 µs gate delay, and 100 µs gate width in the range 200nm to 975nm (resolving power $\lambda/\Delta\lambda \simeq 4 \times 10^3$ measured at 589.60 nm on a Na peak). The irradiance on the sample surface was 190 GW cm^{-2}.

Data on alloys were collected using a Nd:YAG laser (*Quantel Ultra*) operating at a wavelength of 266 nm, 6 mJ pulse energy, and 4 ns pulse length. The plasma emission was analyzed with a *LTB Aryelle 400* spectrometer with a fixed aperture of 50 µm × 50 µm, equipped with an *Andor* DH740 ICCD, in the range 310 nm to 613 nm (resolving power: $\lambda/\Delta\lambda \simeq 1.8\times10^4$). We used a gate delay of 1 µs and a gate width of 0.5 µs. For each sample, 25 spectra were collected, accumulating 20 laser shots for per crater. The irradiance on the surface of the samples was 76 GW cm^{-2}.

V. Training Methodology

Before entering the discussion of the results, we detail some training techniques used to fit the MT model to the LIBS data.

A. Data Curation

For the analysis of the alloy matrices, we use the intensities of the most intense persistent lines of Fe in the spectral range considered, integrated over 10 wavelength channels, as secondary outputs of the MT model. Specifically, we choose the 8 strongest persistent lines, reported in [21]. Given the small number of samples available, 30% of the spectra for each sample is retained as independent test set, while the rest is used as training set for the baseline models, and as input of the data augmentation for the MT architecture. This ensures, on average, an in-sample inference for the algorithms (the selection of test samples comes from the same samples in the training set). Moreover, it provides the means to verify whether the augmentation technique correctly enhances the training distribution. Notice that this does not automatically translate into a simpler task for the model: spectra in the test set may still differ from the training distribution, due to random and local fluctuations, thus they may represent out-of-distribution data on a spectrum basis.

For the cement samples, we consider two molecular bands centred at 593.46 nm and 617.74 nm, and integrated over 14 channels, as secondary outputs. As a general guideline,

choosing many secondary outputs helps the convergence of the network, and preserves a good generalization performance of the model. Samples are separated into training and test sets on a sample basis: 12 matrices are considered as calibration set, while 7 samples are used for inference. In the test set, we insert specifically samples which present a different matrix (type of cement) or manufacturing procedure (different salts added in the mixture) in order to check the out-of-distribution generalization ability of the algorithms, and their ability to recognize possible anomalies.

As preprocessing, outliers are removed from the training set, either experimental or synthetic: at a given wavelength, we define outliers as spectra presenting an intensity outside the interval between the 5th and 95th percentile of the values. The goal is to build performing calibration models without using extreme configurations. In order to test the generalization ability of the models, we retain the outliers in the test set. For the alloy samples, we focus on the Fe line at 373.49 nm (the most intense persistent line), while, for the cement matrices, we consider the molecular band of CaCl at 593.46 nm.

Spectra are normalized using the integral intensity at a given wavelength. Since the procedure is performed independently on each spectrum, we can also safely normalize the spectra in the test sets. For the alloy matrices, we consider the integrated intensity of the most intense emission line of the matrix itself over an interval of 10 wavelength channels. The cement samples have been normalized to the intensity of the CaO molecular band at 615.03 nm, integrated over 20 wavelength channels.

B. Data Analysis and Methodology

We compare our results with several algorithms known in the LIBS literature, namely the classical Linear Regression (LR), MVA Linear Regression (MLR), simple Fully Connected NNs (FCNNs), and PLS1. For LR and MLR, no additional validation sets have been considered, as no free parameters are present in the algorithms. For PLS1, we perform a 5-fold cross-validation procedure. In the case of NNs, we use a single holdout validation set made of 20% of the spectra contained in the training set, selected using a stratified strategy to preserve the fraction of spectra for each sample. Hyperparameters are optimized using a tree-structured Parzen estimator [22]. In order to avoid any data leakage, we avoid using the experimental spectra when optimizing the model. We use common regression metrics such as the Root Mean Squared Error (RMSE), the Mean Absolute Error (MAE), and the MAPE to score the results of the algorithms. Results always refer to the independent experimental test set.

In the case of LR, we consider the integral intensity of selected wavelength channels as inputs of the model. Specifically, we consider the integral intensity of the Fe emission line at 373.49 nm, integrated over 10 wavelength channels, for the alloy matrices. For the cement samples, we use the CaCl molecular band at 593.46 nm, integrated over 14 channels. For the MLR and the FCNN, we select several atomic Fe emission lines as inputs, in the case of the alloy matrices. We use the CaCl molecular bands at 593.46 nm and 617.74 nm for

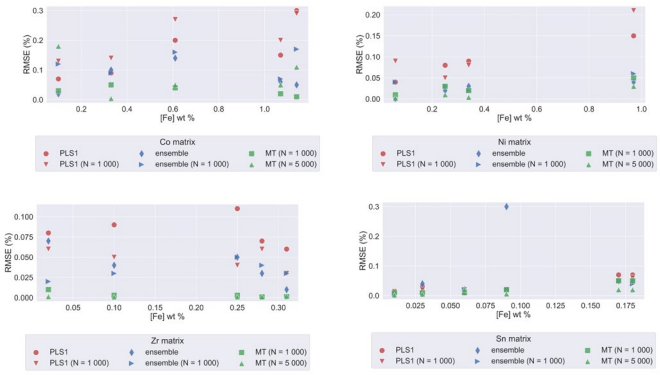

Fig. 6. Prediction uncertainties of the MT model on alloy matrices as a function of training set size.

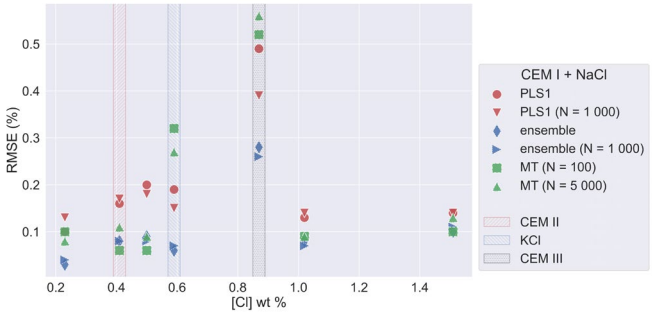

Fig. 7. Prediction uncertainties of the MT model on cement samples as a function of training set size.

the cement samples. On the other hand, for PLS1 and the MT network, we use the entire spectra as input. Finally, we choose a quadratic regression model in the case of LR (without interaction terms), while we consider a simple regression model for MLR.

VI. Results and Discussion

In the analysis, we consider different types of baselines, in order to provide a complete comparison of the proposed technique with the SOTA. Specifically, we show the results of PLS1, which does not require the extraction of information from the experimental spectra, and an ensemble model, with initial feature (emission lines) selection, represented by the best result of LR, MLR, and FCNNs. Our proposed MT model does not require expertise in selecting input data, since the important variables are learnt during training. It requires a degree of knowledge of the emission lines of the analyte to choose the secondary outputs. A simple choice is to consider the full set of persistent lines of the analyte in the available spectral rang, since spectral interference on single lines (or bands) is taken care of by the CNN architecture.

A. Performance of the Model

We consider different aspects relating to the nature of the analysis, such as the dependence on the size of the synthetic training set and the choice of the random noise: we train different models, generating a different number of synthetic spectra, with different β parameters, for each sample. The results with a dependence on the synthetic training set are

graphically summarized in Fig. 6 for the alloy matrices, and in Fig. 7 for the cement samples. A similar behavior has been observed for the dependence of the noise parameter, with the best results appearing for $\beta = 0.10$ in the case of alloy matrices and $\beta = 0.03$ for the cement samples, as expected from the discussion in Section 3.1.

In terms of robustness, the MT architecture is capable of delivering a homoscedastic performance for all concentrations of the analyte (apart from anomalies, which are discussed in the following). Moreover, the prediction uncertainties are usually comparable with or better than the ensemble model, which already presents good results. The MT architecture is capable of selecting the information contained in the data to base its own predictions. The use of synthetic samples enables to capture the fluctuations at low concentrations, where the PLS1 model struggles to give accurate results. As a general remark, the optimal number of synthetic spectra varies for each matrix, depending on the degree of spectral interference, noise, and sparsity (e.g. 5000 spectra per sample in the case of the Zr matrix, or 1000 spectra per sample for the Cu matrix).

Heuristically, we noticed that the creation of many synthetic spectra impacts on training time for less than 1% of the total training time, while the latter grows linearly. Such behavior makes it usually possible to experiment with a few options, in order to determine, using the validation set or the experimental training samples, the best trade-off between the performance of the model and the computational power available.

B. Validation of the Predictions

For the analysis of the trustworthiness of the model, we focus specifically on the predictions of the MT model on the independent test set of the cement samples. In this analysis, ground truth values of the concentration of the analyte are available for a direct comparison with the predictions of the model. However, in a field application, reference values would not be available. The analysis of the secondary outputs of the MT architecture is a tool to assess the confidence of the predictions and detection of the anomalous samples or modifications in the experimental conditions.

We choose two CaCl molecular bands for the analysis of the trustworthiness of the predictions. This represents an easy choice, as the two bands are the two most intense in the spectral range considered. Following previous sections, we compute the confidence intervals on the experimental training data (unseen by the model, which is trained on synthetic spectra). We consider these values as reference in the analysis of the trustworthiness of the model, since they represent known standards, whose labels are available. *A posteriori*, we notice that the predictions of the MT architecture are all compatible with the respective ground truths, even though some samples present larger uncertainties, which may indicate faulty values. However, in the absence of reference values, predictions alone are not sufficient to measure the trustworthiness of the model.

The predictions of the secondary values show that the band at 593.46 nm displays a pattern which identifies some anomalies in the prediction of the integral intensities (see Table 1). As previously shown (see Fig. 3 for a reference), this pattern of MAPE and t-value is typical of anomalous samples, for which the model does not provide precise predictions. In hindsight, the analysis of the secondary outputs identifies the three out-of-distribution samples present in the dataset (different matrix and salt).

TABLE I
PREDICTIONS ON CEMENT SAMPLES

GROUND TRUTH	PREDICTION	MAPE		T-VALUE	
[Cl] wt %	[Cl] wt %	593.46 nm	617.74 nm	593.46 nm	617.74 nm
0.23	0.25 ± 0.08	1.4	0.7	0.69	0.34
0.41 (CEM II)	0.41 ± 0.11	**4.1**	0.8	0.98	0.32
0.50	0.54 ± 0.09	1.7	0.7	0.48	0.26
0.59 (KCl)	0.33 ± 0.27	**4.2**	0.8	1.29	0.31
0.87 (CEM III)	1.40 ± 0.56	**4.7**	**1.4**	0.77	0.41
1.02	0.97 ± 0.09	1.3	0.4	0.38	0.16
1.51	1.54 ± 0.13	1.5	0.8	0.43	0.30

Trustworthiness of the predictions on cement samples.

To measure the trustworthiness of the predictions, we then use a standard Student's two-tailed test (confidence $1 - \alpha = 0.95$ and 25 degrees of freedom) to assess the predictions of the molecular emission bands. We notice that, although the samples register as anomalies, the variance of the predictions is such to include the true values of the secondary outputs in the error intervals with good confidence (case 4 in Fig. 3). Given the interdependencies of the MT model previously discussed, the confidence on the predictions of the secondary output influences directly the confidence on the main prediction, the concentration of the analyte. The precise quantification of the confidence level is, nonetheless, not trivial because of the large number of parameters involved in the computation. In the case at hand, we can interpret the result by noting that the MT model is still capable of providing trustworthy predictions on the concentration of the analyte. However, its precision is highly affected in the presence of anomalous samples: the predictions of the main output contain the true value, within the uncertainty, with high probability. In other words, in this scenario, the predicted values of the concentration of the analyte can be considered compatible with the reference values, provided by the supplier. Further investigation on three anomalous samples remains necessary.

The procedure provides a way to assess the robustness of the MT architecture on the entire range of variability of the Fe concentration. Notice that the analysis of the secondary outputs is quantitative: for instance, a p-value of the prediction can be computed. Moreover, different choices of the confidence $1 - \alpha$ allow comparing the model as functions of the confidence level. Finally, this analysis is always possible, with any sample, as the information contained in the secondary outputs of the model is directly comparable with the experimental data. This is different from the concentration of the analyte itself, which is known only for standard samples.

C. Calibration Transfer and Anomaly Detection

We test the trustworthiness of the predictions in the presence of a change in the distribution of the samples. Specifically, we

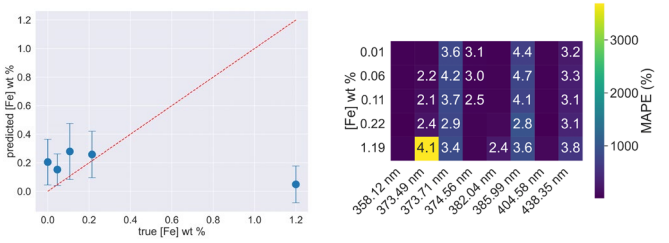

Fig. 8. Comparison of ground truth and predicted values on the Ti matrix, using a model trained on the Co matrix.

consider the MT model trained on the Co matrix, and we perform the inference on the Ti and Zn matrices. In Fig. 8, we show the true and predicted concentrations of the analyte. The a posteriori analysis shows that the model does not generalize among different matrices, hence it is not usable for calibration transfer. However, a real-time application would not reveal the same behavior, as ground truth values would not be available. The Ti matrix presents predictions characterized by large values of the error fractions and by incompatible values of the t variable, whose threshold $\hat{t}^{\nu}_{1-\frac{\alpha}{2}} = 2.06$ for $\nu = 25$ degrees of freedom and a confidence $1 - \alpha = 0.95$. This shows that the predictions of the concentration levels 0.11 wt% and 0.22 wt%, though compatible with their respective ground truths, should not be deemed trustworthy. The MT model is thus capable of detecting anomalies or modifications in the experimental conditions, which is key to assessing correctly the ability of the model to provide trustworthy predictions.

VII. Conclusions

In this work, we use DL techniques to address the quantitative analysis of LIBS data, the prediction of the concentration of an analyte, using a MVA calibration procedure. We focus on complementary aspects: the creation of a synthetic set of spectra as a data augmentation technique to increase the number of samples available for training, the construction of a robust MT learning model based on deep CNNs, and the analysis of the confidence of the predictions. We use the entire experimental emission spectra as inputs, without the need for a preselection of variables or dimensionality reduction. We leverage the robustness and performance of CNNs with the possibility to provide a tool to assess the trustworthiness of the predictions of the model, even for unknown data.

To this end, we introduce a MT learning architecture. The model is capable of predicting the concentration of the analyte and the integral intensities of relevant emission lines (or molecular bands), at the same time. Given the size and complexity of the DL model, we introduce a data simulation technique, to create an arbitrary number of input spectra, statistically representative of the experimental data. The MT architectures display robustness across the range of variation of the analytes. The presence of the secondary outputs, allows us to introduce a statistical analysis, based on the mutual dependencies of the parameters of the AI architecture, which enables the assessment of the trustworthiness of the model. Comparisons of the predicted values with the intensities found in the experimental spectra can be used to study the predictions of the concentration of the analyte, at a given level of confidence of the model. In turn, this grants the ability to assess the extrapolation abilities of the DL model.

Acknowledgements

We acknowledge the financial support of the *Cross-Disciplinary Program on Instrumentation and Detection of CEA*, the French *Alternative Energies and Atomic Energy Commission*. We thank G. Gallou for proposing this collaboration. This publication was made possible by the use of the *FactoryIA* supercomputer, financially supported by the *Ile-De-France Regional Council*.

References

[1] S. Moncayo, L. Duponchel, N. Mousavipak, *et al*., "Exploration of megapixel hyperspectral LIBS images using principal component analysis," *J. Anal. At. Spectrom.*, vol. 33, 210–220, 2 2018. doi: 10.1039/c7ja00398f.

[2] R. Finotello, M. Tamaazousti, and J.-B. Sirven, "HyperPCA: A powerful tool to extract elemental maps from noisy data obtained in LIBS mapping of materials," *Spectrochim. Acta Part B*, vol. 192, 106418, 2022. doi: 10.1016/j.sab.2022.106418.

[3] R. Sattmann, I. Monch, H. Krause, *et al*., "Laser-induced breakdown spectroscopy for polymer identification," *Appl. Spectrosc*., no. 3, 456–461, 1998. doi: 10.1366/0003702981943680.

[4] L.-N. Li, X.-F. Liu, F. Yang, *et al*., "A review of artificial neural network based chemometrics applied in laser-induced breakdown spectroscopy analysis," *Spectrochim. Acta Part B,* vol. 180, 106183, 2021. doi: https://doi.org/10.1016/j. sab.2021.106183.

[5] V. C. Costa, D. V. Babos, J. P. Castro, *et al*., "Calibration strategies applied to laser-induced breakdown spectroscopy: A critical review of advances and challenges," *J. Braz. Chem. Soc.*, vol. 31, no. 12, 2439–2451, 2020. doi: 10.21577/0103-5053.20200175.

[6] V. Motto-Ros, S. Moncayo, F. Trichard, *et al.*, "Investigation of signal extraction in the frame of laser induced breakdown spectroscopy imaging," *Spectrochim. Acta Part B*, vol. 155, 127–133, 2019. doi: 10.1016/j.sab.2019.04.004.

[7] N. C. Dingari, I. Barman, A. K. Myakalwar, et al., "Incorporation of support vector machines in the LIBS toolbox for sensitive and robust classification amidst unexpected sample and system variability," *Anal. Chem*., vol. 84, no. 6, 2686–2694, 2012. doi: 10.1021/ac202755e.

[8] P. Yaroshchyk, D. L. Death, and S. J. Spencer, "Comparison of principal components regression, partial least squares regression, multi-block partial least squares regression, and serial partial least squares regression algorithms for the analysis of Fe in iron ore using LIBS," *J. Anal. At. Spectrom.*, vol. 27, no. 1, 92–98, 2012. doi: 10.1039/c1ja10164a.

[9] T. Takahashi and B. Thornton, "Quantitative methods for compensation of matrix effects and self-absorption in laser induced breakdown spectroscopy signals of solids," *Spectrochim. Acta Part B*, vol. 138, 31–42, 2017. doi: 10.1016/j.sab.2017.09.010.

[10] E. D'Andrea, S. Pagnotta, E. Grifoni, *et al*., "A hybrid calibrationfree/artificial neural networks approach to the quantitative analysis of LIBS spectra," *Appl. Phys. B*, vol. 118, no. 3, 353–360, 2015. doi: 10.1007/s00340-014-5990-z.

[11] L. Narlagiri and V. R. Soma, "Simultaneous quantification of Au and Ag composition from Au-Ag bi-metallic LIBS spectra combined with shallow neural network model for multi-output regression," *Appl. Phys. B: Lasers Opt*., vol. 127, no. 9, 135, 2021. doi: 10.1007/s00340-021-07681-y.

[12] T. Chen, L. Sun, H. Yu, *et al*., "Deep learning with laserinduced breakdown spectroscopy (LIBS) for the classification of rocks based on elemental imaging," *Appl. Geochem*., vol. 136, 105135, 2022. doi: 10.1016/j.apgeochem.2021.105135.

[13] R. Caruana, "Multitask learning: A knowledge-based source of inductive bias," in Proceedings of the Tenth International Conference on International Conference on Machine Learning, ser. Icml'93, Amherst,

MA, USA: Morgan Kaufmann Publishers Inc., 1993, 41–48, isbn: 1558603077. doi: 10.1016 /b978-1-55860-307-3.50012-5.

[14] R. B. Anderson, J. F. Bell III, R. C. Wiens, *et al*., "Clustering and training set selection methods for improving the accuracy of quantitative laser induced breakdown spectroscopy*," Spectrochim. Acta Part B*, vol. 70, 24–32, 2012. doi: 10.1016/j. sab.2012.04.004.

[15] F. O. Borges, G. H. Cavalcanti, G. C. Gomes, *et al*., "A fast method for the calculation of electron number density and temperature in laser-induced breakdown spectroscopy plasmas using artificial neural networks," *Appl. Phys. B*, vol. 117, no. 1, 437–444, 2014. doi: 10.1007/s00340-014-5852-8.

[16] S. Chen, H. Pei, J. Pisonero, *et al*., "Simultaneous determination of lithology and major elements in rocks using laserinduced breakdown spectroscopy (LIBS) coupled with a deep convolutional neural network," *J. Anal. At. Spectrom*., vol. 37, no. 3, 508–516, 2022. doi: 10.1039/d1ja00406a.

[17] J.-M. Mermet, "Limit of quantitation in atomic spectrometry: An unambiguous concept?" *Spectrochim. Acta Part B*, vol. 63, no. 2, 166–182, 2008, Honoring Issue A Collection of Papers on Atomic, Molecular and Laser Spectroscopy Dedicated to James D. Winefordner. doi: 10.1016/j.sab.2007.11.029.

[18] W. Zhao, C. Li, C. Yan, *et al*., "Interpretable deep learning assisted laser-induced breakdown spectroscopy for brand classification of iron ores," Anal. Chim. Acta, vol. 1166, 338574, 2021. doi: 10.1016/j.aca.2021.338574.

[19] X. Zhang, J. Xu, J. Yang, *et al*., "Understanding the learning mechanism of convolutional neural networks in spectral analysis," *Anal. Chim. Acta*, vol. 1119, 41–51, 2020. doi: 10.1016/j.aca.2020.03.055.

[20] T. Völker, G. Wilsch, I. B. Gornushkin, *et al*., "Interlaboratory comparison for quantitative chlorine analysis in cement pastes with laser-induced breakdown spectroscopy," *Spectrochim. Acta Part B*, vol. 202, 106632, 2022. doi: 10.1016/j.sab.2023.106632.

[21] J. Sansonetti, Handbook of Basic Atomic Spectroscopic Data, NIST Standard Reference Database 108, 2003. doi: 10.18434/T4FW23.

[22] J. Bergstra, R. Bardenet, Y. Bengio, *et al*., "Algorithms for hyper-parameter optimization," in Advances in Neural Information Processing Systems, J. Shawe-Taylor, R. Zemel, P. Bartlett, et al., Eds., vol. 24, Curran Associates, Inc., 2011, doi:10.5555/2986459.2986743.