



OPEN

DATA DESCRIPTOR

# Benchmark classification dataset for laser-induced breakdown spectroscopy

Erik Képeš<sup>1,2</sup> , Jakub Vrábel<sup>1,2</sup> , Sára Střítežská<sup>1</sup>, Pavel Pořízka<sup>1</sup> & Jozef Kaiser<sup>1</sup> 

In this work, we present an extensive dataset of laser-induced breakdown spectroscopy (LIBS) spectra for the pre-training and evaluation of LIBS classification models. LIBS is a well-established spectroscopic method for *in-situ* and industrial applications, where LIBS is primarily applied for clustering and classification tasks. As such, our dataset is aimed at helping with the development and testing of classification and clustering methodologies. Moreover, the dataset could be used to pre-train classification models for applications where the amount of available data is limited. The dataset consists of LIBS spectra of 138 soil samples belonging to 12 distinct classes. The spectra were acquired with a state-of-the-art LIBS system. Lastly, the composition of each sample is also provided, including estimated uncertainties.


## Background & Summary

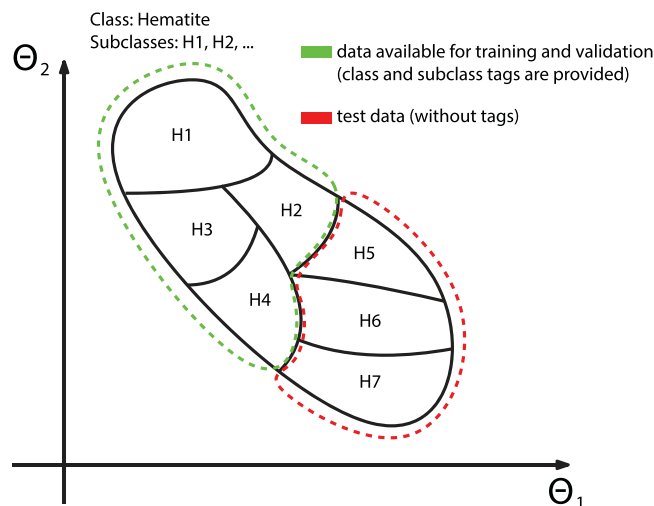
Laser-induced breakdown spectroscopy (LIBS) is an emission spectroscopic method that uses a high-powered laser pulse to ignite a microplasma. This is achieved by focusing laser pulses with lengths in the fs–ns range into spots with diameters of tens of  $\mu\text{m}$ . Consequently, a sufficiently high energy fluence is reached to ionize the target material. Subsequently, emission of the ignited plasma is collected, dispersed by a spectrometer, and recorded. Assuming a stoichiometric ablation, the dispersed light intensities can be related to the composition of the target material<sup>1,2</sup>.

Owing to these relatively simple principles, LIBS instrumentations are generally robust. Consequently, LIBS is often preferred in industrial settings that are unfavourable for most common spectroscopic methods, such as charged-particle-based techniques and variations of mass spectrometry. As such, LIBS has been widely adapted in geology<sup>3,4</sup>, steel industry<sup>5</sup>, and forensics<sup>6</sup>. Nevertheless, recently, LIBS has been gaining a foothold in various biological applications, e.g., mapping of biological samples<sup>7,8</sup>.

Generally, most successful applications of LIBS are clustering and classification<sup>9,10</sup>. Meanwhile, the current limitations of LIBS inhibit LIBS from being reliably applied for quantification. Consequently, there is a relatively wide range of literature reporting on the classification of various materials using LIBS. As such, the classification approaches also vary significantly, including the spectral pre-processing, feature engineering, the classification model itself<sup>11</sup>. Hence, the systematic comparison of the various approaches is not possible. Moreover, a common approach to classification is the randomized division of the complete dataset into training, validation, and testing subsets. Hence, this approach relies on the testing dataset comprising emission spectra that were collected during the same measurement as the training spectra. However, in practical applications, fresh data is constantly being evaluated by the existing model, i.e., the testing dataset is constantly evolving.

Consequently–inspired by the recent breakthroughs in image recognition tasks partially made possible by datasets such as MNIST handwritten digit dataset<sup>12</sup>–we propose a similar dataset for LIBS. The dataset is constructed from geological samples, where several distinct samples belong to the same class. Thus, we propose a dataset where the training and testing data is sampled from distinct materials. As such, classifiers that perform well on the proposed dataset must be able to generalize rather than simply learn the distribution of the data. The classification problem is shown schematically in Fig. 1: Various ore samples belong to the same geological class, e.g., the class hematite is represented by six samples. Four of these samples are provided in the training dataset, while the remaining two are included in the test dataset. Considering the interclass variability, this classification

<sup>1</sup>Central European Institute of Technology, Brno University of Technology, Purkyňova 123, 612 00, Brno, Czech Republic. <sup>2</sup>These authors contributed equally: Erik Képeš and Jakub Vrábel.  e-mail: [erik.kepes@ceitec.vutbr.cz](mailto:erik.kepes@ceitec.vutbr.cz)



**Fig. 1** Schematic representation of the classification task: samples belonging to the geological class of hematite in an arbitrary two-dimensional feature space.

task is more challenging than the generally reported cases since the model is expected to generalize from the samples in the green region in order to accurately classify the samples in the red region.

The dataset is expected to provide not only a benchmark for LIBS classification models but also the means of testing classification models for robustness or to identify overtraining. Moreover, the dataset could prove useful for implementing transfer learning for applications where only relatively small datasets are available. Lastly, the dataset could be used to develop feature-engineering and dimensionality-reduction methodologies.

## Methods

**Sample preparation.** The samples comprised certified reference materials (soils) purchased from Ore Research & Exploration Pty Ltd (Melbourne, Australia) and dental gypsum (Spofadental, Czechia) in a weight ratio of 1:1, i.e., each sample consisted of 50 wt.% certified reference soil powder (see below) and 50 wt.% gypsum powder. The samples were prepared by mixing 400 mg of dry soil powder and 400 mg of dry gypsum powder. The weight of the constituents was measured with a TP-303 laboratory scale (Denver Instruments, Germany) with an instrumental uncertainty of  $\pm 1$  mg. In total, 46 soil samples were used (table included in the data repository). Each of the 46 soil powders belongs to one of 12 ore types. The latter classification was provided by the vendor and can be found in the *OREAS* tab of the *support\_tables.xlsx* excel file accessible in the data repository. For simplicity, in the upcoming description of the sample preparation, the gypsum portion of the samples is considered to be constant (within the measurement uncertainty) and is excluded.

To adjust the classification difficulty of the dataset in a controlled manner, the soil powders were mixed to obtain a certain degree of intraclass similarity: Each of the 46 soil samples (base sample) was mixed with two soil powders (additive samples) from a different ore class. Consequently, 138 samples were obtained in total. During the mixing,  $\frac{1}{4}$  of the base sample's mass was replaced with one of the additives. Subsequently, 0.5 ml water was added followed by another mixing. Lastly, the wet mixture was poured into a small plastic container, where it formed a flat surface.

Although the dataset is not meant for quantitative analysis, the composition of the samples is provided in the data repository including estimated uncertainties. The uncertainties are provided without considering the influence of the added water.

**Measurements.** The samples were measured in a state-of-the-art LIBS interaction chamber that enables the precise control of the measurement parameters, including the atmosphere. As such, the highest standards of LIBS measurements were maintained. The samples were mapped with a 100  $\mu\text{m}$  step size (distance between shots) at a 20 Hz ablation repetition rate with a pulse energy of 15 mJ at the ablation wavelength of 532 nm (Nd:YAG, 10 ns pulse length, CFR400, Quantel, France). The ablation crater diameter measured under an optical microscope was 60  $\mu\text{m}$ . The optical emission of the laser-induced plasma was collected using a single lens and guided to the entrance slit of an echelle spectrograph (EMU 65, Catalina Scientific, US; resolving power  $R = 6000$ ) by an optical fibre. The light resolved by the spectrometer (the echellogram) was recorded with an EMCCD camera (Falcon Blue, Raptor Photonic, IR) and translated into spectra using the control software proprietary to the spectrometer (KestrelSpec™ Imaging Spectroscopy, Catalina Scientific, US). The camera recorded the incoming light with a delay of 0.3  $\mu\text{s}$  after the ablation laser pulse (commonly referred to as the gate delay) and for the duration of 50  $\mu\text{s}$  (commonly referred to as the gate width). These timing values have been chosen following an optimization procedure based on the signal-to-baseline ratio, where the height of an arbitrary emission line of a non-matrix element has been considered as the signal. The measurements were carried out in air. Moreover, during the measurements, the sample surface was continuously purged by air with a volumetric flow rate of 10 l/min.

## Data Records

The dataset (available on Figshare<sup>13</sup>) consists of two hdf5 files, a .csv file and a .xlsx table. The first hdf5 file (*train.h5*) contains the training dataset which is advised to be further divided into a training and validation dataset. The training dataset includes class labels. The second hdf5 (*test.h5*) file contains the testing data without class labels. The rows of each dataset correspond to a single emission spectrum obtained by laser-induced breakdown spectroscopy. Consequently, the columns correspond to distinct wavelength values and the elements of the dataset are intensity values in arbitrary units (a.u.), which is a common representation of emission intensity in the LIBS community. The training dataset contains 500 spectra for each sample. Nevertheless, the users are welcome to load in only a subset of the dataset (which is straightforward with the supported code). Meanwhile, the testing dataset contains a varying number of spectra for each sample.

The class labels of the testing dataset are provided in the form of a .csv file titled *test\_labels.csv*. Lastly, the composition of the samples is given in an .xlsx file (*support\_tables.xlsx*). The excel file contains 4 spreadsheets: *OREAS* lists the composition of the certified soil powders as provided by the vendor; *MIXED\_composition* lists the estimated composition of the mixed samples (excluding the gypsum fraction); *MIXED\_uncertainty* lists the estimated uncertainties of the mixed compositions; and *MIXED\_combined* lists the mixed compositions and uncertainties in a more compact form. The aim of providing the composition and the uncertainties in separate tables is to ease their import for data processing. Consequently, the table combining both the composition and uncertainties facilitates the presentation of the compositions.

## Technical Validation

The composition of the samples is provided with relevant uncertainties in the data repository. The soil samples are certified standard materials. Hence, their composition was determined by the vendor. The uncertainty of the constituents' weight fraction ranges from 4 to 10%. However, for a more modest uncertainty estimation, a constant uncertainty of 10% was considered, e.g., the uncertainty of an element present in the soil with a weight fraction of 10 wt.% was  $\pm 1$  wt.%. The final combined uncertainty was determined from the non-linear uncertainty propagation as:

$$\Delta_f = \sqrt{\sum_{i=1}^N \left( \frac{\partial f_e(x_1, \dots, x_N)}{\partial x_i} \Delta(x_i) \right)^2}$$

where  $\Delta_f$  is the combined uncertainty,  $\Delta(x_i)$  is the uncertainty of  $x_i$ ,  $f_e(x_1, \dots, x_N)$  is the function describing the weight fraction of an analyte  $e$  in the final sample; namely:

$$f_e(x_1, \dots, x_N) = W_e = \frac{M_{s,1} \cdot w_{e,1} + M_{s,2} \cdot w_{e,2}}{M_{s,1} + M_{s,2} + M_G}$$

where  $W_e$  is the weight fraction of analyte  $e$  in the sample;  $M_{s,k}$  and  $w_{e,k}$  are the weight of soil sample  $k$  and the analyte's weight fraction in soil  $k$ , respectively; and  $M_G$  is the weight of the added gypsum powder. For samples mixed from a single soil standard,  $M_{s,2} = 0$ . Consequently, the uncertainty of analyte  $e$  in the final sample is given as:

$$\Delta(W_e) = \sqrt{\left( \frac{w_{e,1} \cdot (M_{s,1} + M_{s,2} + M_G) - M_{s,1} \cdot w_{e,1}}{(M_{s,1} + M_{s,2} + M_G)^2} \cdot \Delta(M_{s,1}) \right)^2 + \left( \frac{w_{e,2} \cdot (M_{s,1} + M_{s,2} + M_G) - M_{s,2} \cdot w_{e,2}}{(M_{s,1} + M_{s,2} + M_G)^2} \cdot \Delta(M_{s,2}) \right)^2 + \left( \frac{M_{s,1}}{M_{s,1} + M_{s,2} + M_G} \cdot \Delta(w_{e,1}) \right)^2 + \left( \frac{M_{s,2}}{M_{s,1} + M_{s,2} + M_G} \cdot \Delta(w_{e,2}) \right)^2 + \left( \frac{M_{s,1} \cdot w_{e,1} + M_{s,2} \cdot w_{e,2}}{(M_{s,1} + M_{s,2} + M_G)^2} \cdot \Delta(M_G) \right)^2}$$

where  $\Delta(X)$  is the uncertainty of the quantity  $X$ . Lastly, the uncertainty of the weight fraction of element  $e$  in the soil sample  $k$  is determined as  $\Delta(w_{e,1}) = 0.1 \cdot w_{e,k}$ .

The dataset was classified as part of a competition held at the EMSLIBS19 conference (<http://libs.ceitec.cz/libs-contest/>). The highest accuracy achieved was approximately 90%. Two additional approaches reached classification accuracies over 80%. Details of the applied methodologies will be specified elsewhere. Nevertheless, the classification of the dataset has been proven to be adequately challenging to serve as a benchmark dataset.

## Code availability

Custom code for loading in the training and testing datasets is available in the data repository for Python, R, and MATLAB. The Python code was tested in Python 3.6 and requires the following libraries: “os”, “h5py”, and “numpy”. The R code was tested in R 3.5.2 and requires the following libraries: “rhd5”. Lastly, the MATLAB code was tested in MATLAB 2016. The codes are intended to load in the data from the hdf5 files in a user-friendly manner.

Received: 4 December 2019; Accepted: 29 January 2020;

Published online: 13 February 2020

## References

- Miziolek, A. W., Palleschi, V. & Schechter, I. *Laser-Induced Breakdown Spectroscopy (LIBS): Fundamentals And Applications*. (Cambridge University Press, 2006).
- Noll, R. *Laser-Induced Breakdown Spectroscopy*. (Springer Berlin Heidelberg, 2012).
- Harmon, R. S., Russo, R. E. & Hark, R. R. Applications of laser-induced breakdown spectroscopy for geochemical and environmental analysis: A comprehensive review. *Spectrochim. Acta - Part B At. Spectrosc.* **87**, 11–26 (2013).
- Wiens, R. C. *et al.* The ChemCam instrument suite on the Mars Science Laboratory (MSL) rover: Body unit and combined system tests. *Space Sci. Rev.* **170**, 167–227 (2012).
- Noll, R., Fricke-Begemann, C., Connemann, S., Meinhardt, C. & Sturm, V. LIBS analyses for industrial applications – an overview of developments from 2014 to 2018. *J. Anal. At. Spectrom.* **33**, 945–956 (2018).
- De Giacomo, A., Koral, C., Valenza, G., Gaudiuso, R. & Dell’Aglia, M. Nanoparticle enhanced laser-induced breakdown spectroscopy for microdrop analysis at subppm Level. *Anal. Chem.* **88**, 5251–5257 (2016).
- Jolivet, L. *et al.* Review of the recent advances and applications of LIBS-based imaging. *Spectrochim. Acta - Part B At. Spectrosc.* **151**, 41–53 (2019).
- Gaudiuso, R. *et al.* Laser-induced breakdown spectroscopy for human and animal health: A review. *Spectrochim. Acta - Part B At. Spectrosc.* **152**, 123–148 (2019).
- El Haddad, J., Canioni, L. & Bousquet, B. Good practices in LIBS analysis: Review and advices. *Spectrochim. Acta - Part B At. Spectrosc.* **101**, 171–182 (2014).
- Pořízka, P. *et al.* On the utilization of principal component analysis in laser-induced breakdown spectroscopy data analysis, a review. *Spectrochim. Acta - Part B At. Spectrosc.* **148**, 65–82 (2018).
- Motto-Ros, V. *et al.* Critical aspects of data analysis for quantification in laser-induced breakdown spectroscopy. *Spectrochim. Acta - Part B At. Spectrosc.* **140**, 54–64 (2018).
- LeCun, Y., Bottou, L., Bengio, Y. & Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **86**, 2278–2323 (1998).
- Képeš, E., Vrabel, J., Strítežská, S., Pořízka, P. & Kaiser, J. Benchmark classification dataset for laser-induced breakdown spectroscopy. *figshare*. <https://doi.org/10.6084/m9.figshare.c.4768790> (2020).

## Acknowledgements

The creation of this dataset has been supported by the Ministry of Education, Youth and Sports of the Czech Republic under the project CEITEC 2020 (LQ1601) and by the CEITEC Nano Research Infrastructure (MEYS CR, 2016–2019), CEITEC Nano + project, ID CZ.02.1.01/0.0/0.0/16\_013/0001728. EK is grateful for the support provided by the grant CEITEC VUT-J-19-5998 from the Brno University of Technology.

## Author contributions

E.K. and J.V. devised the dataset, i.e., the samples, their preparation, and their measurement. P.P. provided advice during the design of the dataset. S.S. carried out the sample preparation and the measurements. J.K. supervised the work to ensure its scientific integrity.

## Competing interests

The authors declare no competing interests.

## Additional information

**Correspondence** and requests for materials should be addressed to E.K.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher’s note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

The Creative Commons Public Domain Dedication waiver <http://creativecommons.org/publicdomain/zero/1.0/> applies to the metadata files associated with this article.

© The Author(s) 2020