Revalue Geoscientific Data Utilising Deep Learning

Joël Morgenthaler[1], Dominik Frefel[2], Joshua Meier[2],
Nils Oesterling[3], Gérard Perren[1,4], Stefan Heuberger[1]

Keywords: geological archive, artificial intelligence, Optical Character Recognition (OCR), text classification, object detection, Natural Language Processing (NLP), Computer Vision (CV)

Abstract
Today, a huge amount of valuable geoscientific data is unused and stored in analogue and poorly accessible paper archives. We introduce state-of-the-art Natural Language Processing (NLP) and Computer Vision (CV) methods for revaluating these geological archives. Scanned geological documents are made machine-readable using Optical Character Recognition (OCR) and are then classified to predefined geoscientific classes by a natural multi-language model conjoined with a multi-class deep neural network prediction head. Moreover, objects like maps, profiles, well logs or graphics are detected employing an object detection model. Trainings on both models were performed on relatively small training sets. However, by optimising the hyperparameter space utilising random search, we find excellent scores for the optimised models benchmarking today's capabilities. As a result, we provide trained models on data from the Swiss Geological Survey (SGS), which are freely available and can be used to revaluate other geological datasets.

Zusammenfassung
Heute liegt eine grosse Menge wertvoller geowissenschaftlicher Daten ungenutzt in analogen und schlecht zugänglichen Papierarchiven. Mit unserem Projekt stellen wir mit Methoden aus dem Gebiet «Natural Language Processing» (NLP) und «Computer Vision» (CV) Möglichkeiten vor, um diese geologischen Archive aufzuwerten. Gescannte geologische Dokumente werden mittels Texterkennung (OCR) maschinenlesbar gemacht und anschliessend durch ein Mehrsprachenmodell in Kombination mit einem Mehrklassen-Vorhersagekopf eines tiefen neuronalen Netzes vordefinierten, geowissenschaftlichen Klassen zugeordnet. Zusätzlich werden Objekte wie Karten, Profile, Bohrprofile oder Grafiken durch ein Objekterkennungsmodell erkannt. Beide Modelle wurden mit kleinen Trainingssätzen trainiert. Durch die Optimierung der Modellparameter mittels Zufallssuche konnten wir jedoch hervorragende Ergebnisse für die optimierten Modelle erzielen, welche den heutigen Möglichkeiten entsprechen. Als Ergebnis stellen wir Modelle basierend auf Daten der Landesgeologie zur Verfügung, welche frei verfügbar sind und für die Prozessierung anderer geologischer Datensätze und Archive verwendet werden können.

1 Introduction

During the last decade, digitalisation has made great progress in many business processes of various fields. In the field of geology, the digital transformation is in most cases only slowly arising and the opportunities for analysing ancient mostly analogue geological data by employing artificial intelligence is huge. In Switzerland, the Federal Council has given an important impetus towards digital transformation within the geoscience field with the approval of the action plan «Digitisation of the subsurface» (swisstopo 2021) which stays in relation to the report on measures originated from the «Postulate Vogler, 16.4108: Geological data on the subsurface» (Swiss Council 2018). Three of the eight measures of the action plan aim to advance the digitisation of geological archives and to facilitate the accessibility and usabil-

1 Georesources Switzerland Group, Department of Earth Sciences, ETH Zürich, Sonneggstrasse 5, 8092 Zürich
2 Institute for Data Science, FHNW University of Applied Sciences and Arts Northwestern Switzerland, Bahnhofstrasse 6, 5210 Windisch
3 Swiss Geological Survey, Swiss Federal Office of Topography swisstopo, Seftigenstrasse 264, 3084 Wabern
4 Microscopy and Magnetism Group, Photon Science Division, Paul Scherrer Institute, Forschungsstrasse 111, 5232 Villigen

ity of geological data. In this context, digitised means scanned, fully text-recognised and further processed by machine learning algorithms for full use of the original analogue geological data.

Starting with the geoscientific archive of the Swiss Geological Survey (SGS), we develop methods to revaluate the already scanned geoscientific reports. The archive is currently managed with a meta-database storing attributes like title, author, location, and data usage restrictions (Hayoz et al. 2009). Today, it is a time-consuming and inefficient task to find the most appropriate document for a given query as a content-related description as well as full-text search is missing. Therefore, the SGS wants to build a new semi-automated data management workflow, which accounts for the introduced digitalisation mission. Hence, applications are sought for (a) easy exchange of freely accessible geological data between projects and other authorities or service companies, (b) filtering geological data with customised queries and full-text search and (c) providing detailed information on instances such as figures within geological reports. With our current project, we provide methods for filling parts of this gap by automatically assigning each geological asset to a predefined geological class and by detecting graphical objects. We engineer and optimise state-of-the-art Natural Language Processing (NLP) and Computer Vision (CV) models leading to two pipelines for text classification and object detection. Moreover, a web application is designed to interact with the underlying database to pool all available metadata and capitalise on the NLP and CV predictions.

The following study is based on SGS-specific use cases. Hence, data and metadata are selected and provided by the SGS. However, all methods and processing steps can be applied to any geoscientific archive or repository containing scanned documents. The source code is deployed and freely available under the following Git repository: https://gitlab.ethz.ch/swisstopo/revaluatearchive

2 Materials and Methods

The Swiss Geological Archive

The SGS maintains a geological archive containing approximately 45'000 scanned geoscientific assets such as geological reports, drilling profiles, maps and other documents from geological investigations in Switzerland. Figure 1b reveals a simplification of the current data model of the SGS internal management system, which is in operation since 2006 (Hayoz et al. 2009). In total, a set of around 20 attributes are recorded for each asset. Currently, a new data model for geological assets is under development as well as new applications for archiving and querying documents (Oesterling et al. 2022; Brodhag 2022).

Data Labelling & Geoscientific Classes

For the two proposed tasks, text classification and object detection, sample data are provided and annotated by the SGS. For this purpose, the SGS team specified 14 classes for text classification and 10 instance types for the object detection task based on data models, internal applications and domain knowledge (Tab. 1).

The annotation itself is carried out using the computer vision annotation tool CVAT (Sekachev et al. 2019). In particular, the content of each pdf document is first carefully checked by eye and then assigned to one of the predefined geoscientific classes. Similarly, all instances are visually surveyed, accurately framed by ground truth bounding boxes and allocated to the 10 predefined instance types. All classes, instance types, number of documents per class and number of elements per type are summarised in Table 1.

Workflow & Pre-processing

We develop pipelines consisting of three sections with pre-processing, model training and inference for both tasks. Whereas

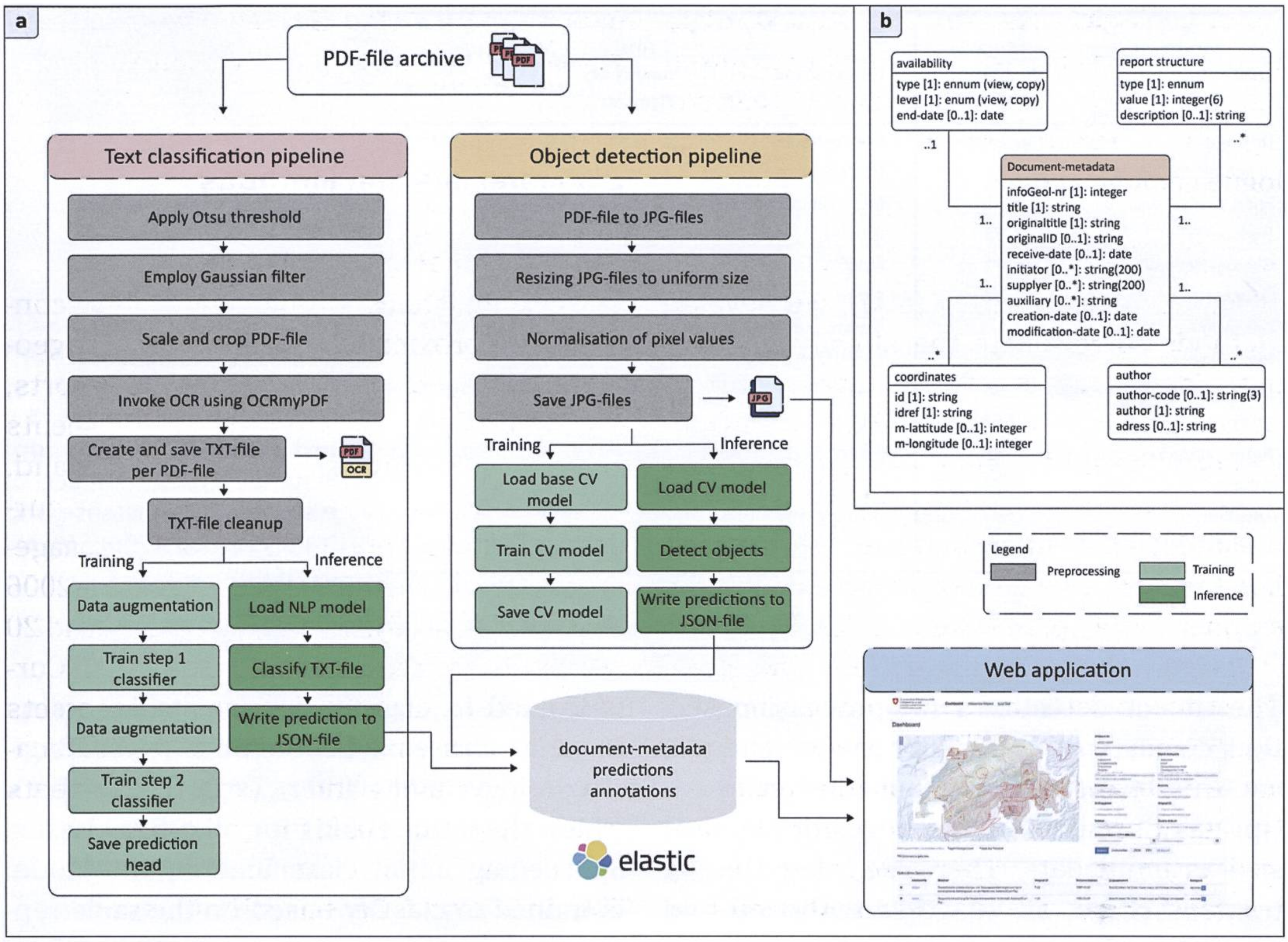

Fig. 1: (a) Processing pipeline and storage workflow with pre-processing (grey), training (light green) and inference (green) for text classification and object detection. Both pipelines access the pdf file archive through a file share. The corresponding predicted classification results are stored in the document metadata (text classification) and predictions (objects), respectively, by utilising the Elasticsearch python API. The web application framework loads the content directly from Elasticsearch and the corresponding file share with all processed files. (b) Modified and simplified data model from the SGS current meta data management system, which is used for the underlying metadata in the web application (modified after Hayoz et al. 2009).

inference denotes the process of predicting classes using the trained model. Figure 1a illustrates these pipelines together with the underlying database and the connection to the designed web frontend.

The pre-processing for the text classification consists basically of the OCR process. However, several processing steps before carrying out the actual OCR need to be done to enhance OCR results. Therefore, preparation steps of the scanned pdf files such as Otsu thresholding, Gaussian filtering for reducing background noise (Davies 2012, p. 40-42), scaling pages to A4 format and border cropping, where artifacts from scanning are carefully removed, need to be executed. Afterwards, the actual OCR is carried out using the open-source application OCRmyPDF in conjunction with the open-source tesseract engine (Kay 2007). Then, the text layer of each analysed pdf is stored as a separate text file. To enhance the model input, modifications of these text files need to be done. Therefore, all symbols except "'?!.,:azAZüöäÜÖÄèéà as well as hyphens and words repeating more than twice are removed. Note that this pre-processing steps without the last modification can be used as a standalone OCR pipeline.

Class	Total Number of documents
Drilling	132
Economic resources	45
Energy resources	45
Geoenergy	83
Geology general	23
Geophysics	74
Geotechnics	178
Hydrology	56
Measurements/Lab results	65
Mining	60
Natural hazards	86
Other	74
Pollution	91
Science	102

Instance type	Total Number of elements	Number in train split	Number in test split
Drilling profile	127	112	15
Geological profile	72	60	12
Geotechnical profile	26	25	1
Graphic	59	56	3
Map	110	99	11
Photo	43	39	4
Seismic profile	33	31	2
Stratigraphic profile	15	14	1
Table	89	82	7
Well Log	102	90	12

Tab. 1: Predefined geoscientific classes: (left) classes and number of documents per class for text classification, (right) instance types, number of elements per instance and train/test split for object detection.

The object detection's pre-processing section consists of four major steps. Therein, each pdf page is converted into an image (jpg) as CV models uses images of identical size as input data. Therefore, after the jpg transformation, all images are resized and normalised to A4 format in order to keep the model input consistent.

Models

For text classification, we employ the open-source FARM framework (Framework for Adapting Representation Models) for training and inference. In essence, FARM facilitates the deployment of NLP models by providing a wide variety of language models and prediction heads that can be combined efficiently and reliably (FARM 2021). Here, we put to work the BERT (Bidirectional Encoder Representations from Transformers) multi-language model together with a text classification multi-class prediction head. BERT is especially designed to use context. As a result, the embedding of a single word can change depending on its usage (Devlin et al. 2019). For training, as illustrated in Figure 1a, we augment the data as the labelled number per geoscientific classes is relatively small (Tab. 1). Moreover, we apply a stepwise classification as the occurrence of the geoscientific classes is imbalanced. First, the model is trained to classify documents according to the classes «Geotechnics», «Drilling», «Hydrology» and «Other» (step 1 classifier), where the latter stands for all other classes. Succeeding initial classification, the model is trained to classify based on the underrepresented but uniformly distributed classes comprising «Other» (step 2 classifier). To compensate for the general data shortage, the model is rigorously trained on several different training and test sets. The ratio of the train/test split always amounts to 0.9/0.1. To maximise the performance, special attention is paid to hyperparameters by optimising them using random search (Bergstra & Bengio 2012). In total, 10 hyperparameters are sampled per iteration.

For object detection, the pre-trained Faster R-CNN/X101-FPN model provided by Detectron2 is applied for training and inference as it reveals highest average precision on the baseline task (Wu et al. 2019). Detectron2, released in 2019, is the second generation of Facebook AI Research's (FAIR) object detection platform and fully implemented in PyTorch (Paszke et al. 2019). As illustrated in Figure 1a, the complete model is saved after training and then reloaded for inference. The number of labelled instances is relatively small (Tab. 1), nevertheless, augmentation is not necessary as it would not improve the results. It is not feasible to average over several different train/test splits as it was done for the text classification due to the model uses high amounts of computational power. The computational power increases linearly with the number of test/train splits. In order to ensure that each labelled sample occurs in the test set, we would need at least five test/train splits. In our case, this would correspond to a computing time of 10 days per iteration instead of two. Therefore, we apply a fixed train/test set with a ratio of approximately 0.9/0.1 (see Tab. 1). For optimisation, six hyperparameters are sampled per iteration and optimised by means of random search (Bergstra & Bengio 2012).

Database & Web Application

As previously mentioned and illustrated in Figure 1a, a web application to pool the gathered results with the already existing meta data is engineered. We use Elasticsearch as an underlying search engine as it provides strong multi-search and full-text search (Gormley & Tong 2015). However, the two machine learning pipelines are completely independent of the database and can be used standalone to process geological assets. Input and output files can be specified manually. The web interface is designed using Django, a high-level open-source Python web framework encouraging rapid development and clean, pragmatic design (Django Software Foundation 2013). For the database architecture, the data model (Fig. 1b) together with new attributes from the machine learning tasks are used. This means, additional attributes for annotated and predicted classes as well as for objects are added. For the latter, additional bounding boxes for each object are required and need to be saved in the database. In our case, the whole application is deployed with docker container (Merkel 2014) on the Amazon Web Service (AWS) account of the Federal Office of Topography swisstopo.

3 Results and Discussion

Text Classification

As introduced, special attention is paid to hyperparameter optimisation. Figure 2a displays the evolution of the F1 scores as a function of iteration length. The F1 scores are based on k-fold cross-validation and hence represent averaged values. With 0.923 and 0.905, we obtain two excellent F1 scores by sampling the hyperparameter space utilising random search. Interestingly, the two classifiers reveal a completely different optimisation behaviour. In particular, the step 2 F1 score is exceptionally bipartite, indicating two distinct sets of hyperparameters. A priori, it seems hardly possible to rationalise this observation. However, differences in the hyperparameters show that the dissimilar number of classes for step 1 and step 2 classifier as well as indistinct geological classes may lead to this pattern. A single classification result for a randomly chosen test set is shown in Figure 2b. The corresponding non-averaged F1 scores are 0.925 (step 1) and 0.857 (step 2), respectively. As evident from the confusion matrix, the document classification model performs remarkably well, compared to today's capabilities. Especially for the step 2 classifier, only a single class, namely «Mining», is moderately more often predicted than labelled. It should, however, be noted that this class is by no means exceptional. The described behaviour must hence be attributed to the chosen test set.

Object Detection

For the object detection model, a mean Average Precision (mAP) of 72.52 is achieved after around 65 iterations by applying random search for optimising the hyperparameter space. Due to high computation demand, the sampling is terminated after 90 iterations although marginally better precisions might have be achieved by iterating for even longer times. Table 2 displays the Average Precision (AP) per instance type

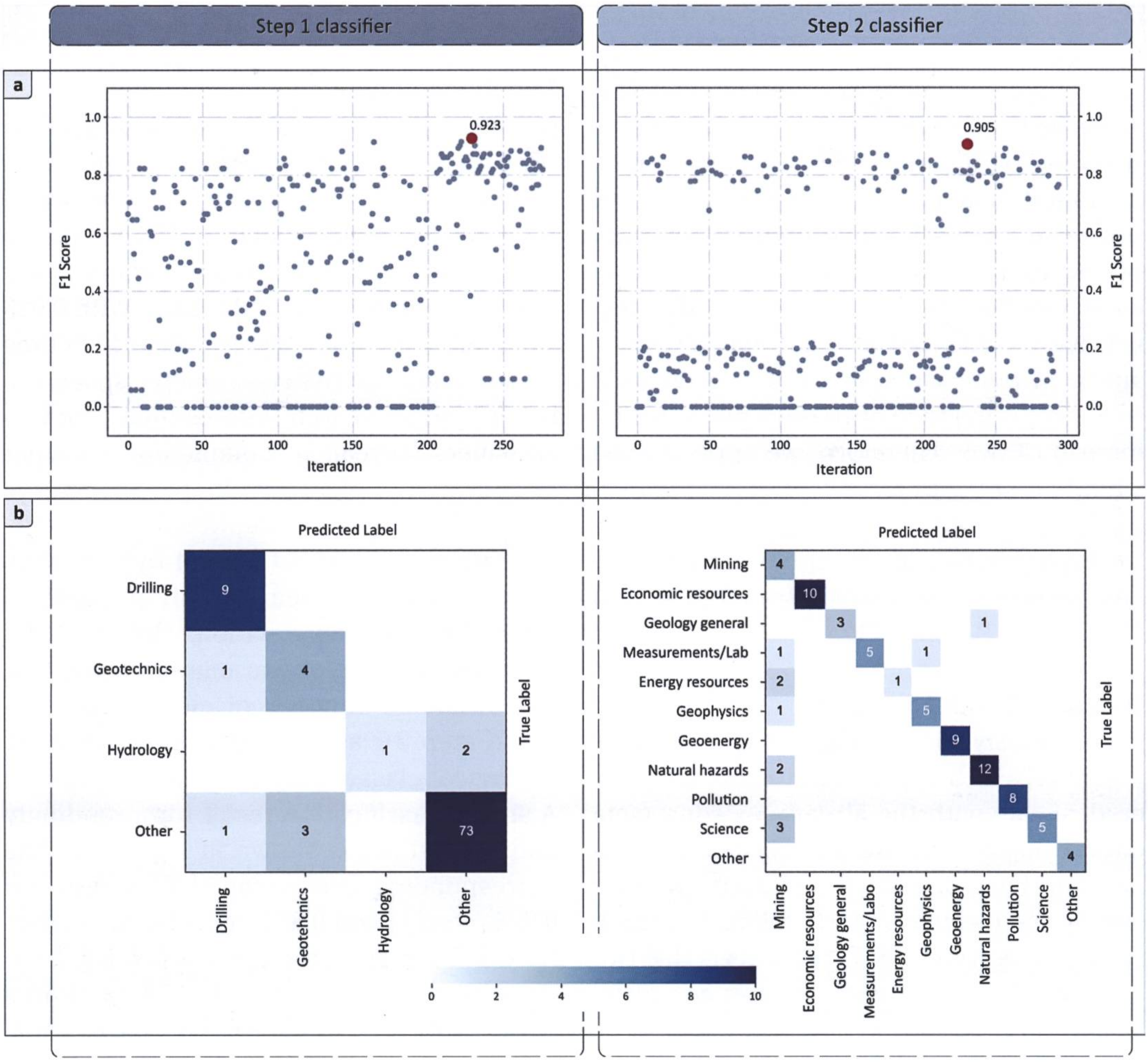

Fig. 2: (a) F1 score as a function of iteration length for both classifiers, step 1 (left) and step 2 (right). The red points highlight the optimal values found by random search. (b) Confusion matrices for step 1 (left) and step 2 (right) classifier based on a randomly chosen test set with non-averaged F1 scores of 0.925 (step 1) and 0.857 (step 2).

with the optimal hyperparameter set. The AP varies from 61.69 up to 86.34. However, three instance types show AP of 100 or 0 and cannot be considered. There is only one or two objects within the corresponding test split as the labelled data is too small. Therefore, an AP of 100 or 0 can be seen only as a random hit. The instance type «Photo» and «Map» achieve the worst APs with 63.35 and 61.67, respectively. For the latter, this can be explained by the fact that maps also can

Instance type	Average Precision (AP)
Drilling profile	80.81
Geological profile	86.34
Geotechnical profile	100
Graphic	75.64
Map	61.67
Photo	65.35
Seismic profile	100
Stratigraphic profile	0
Table	71.63
Well-Log	83.95
AP50	77.39
AP75	73.96

Tab 2: Average Precision per instance for optimal hyperparameter set with a mean average precision of 72.54 and an AP50 of 77.39 (i.e., AP with IoU › 0.5) and an AP75 of 73.96 (i.e., AP with IoU › 0.75).

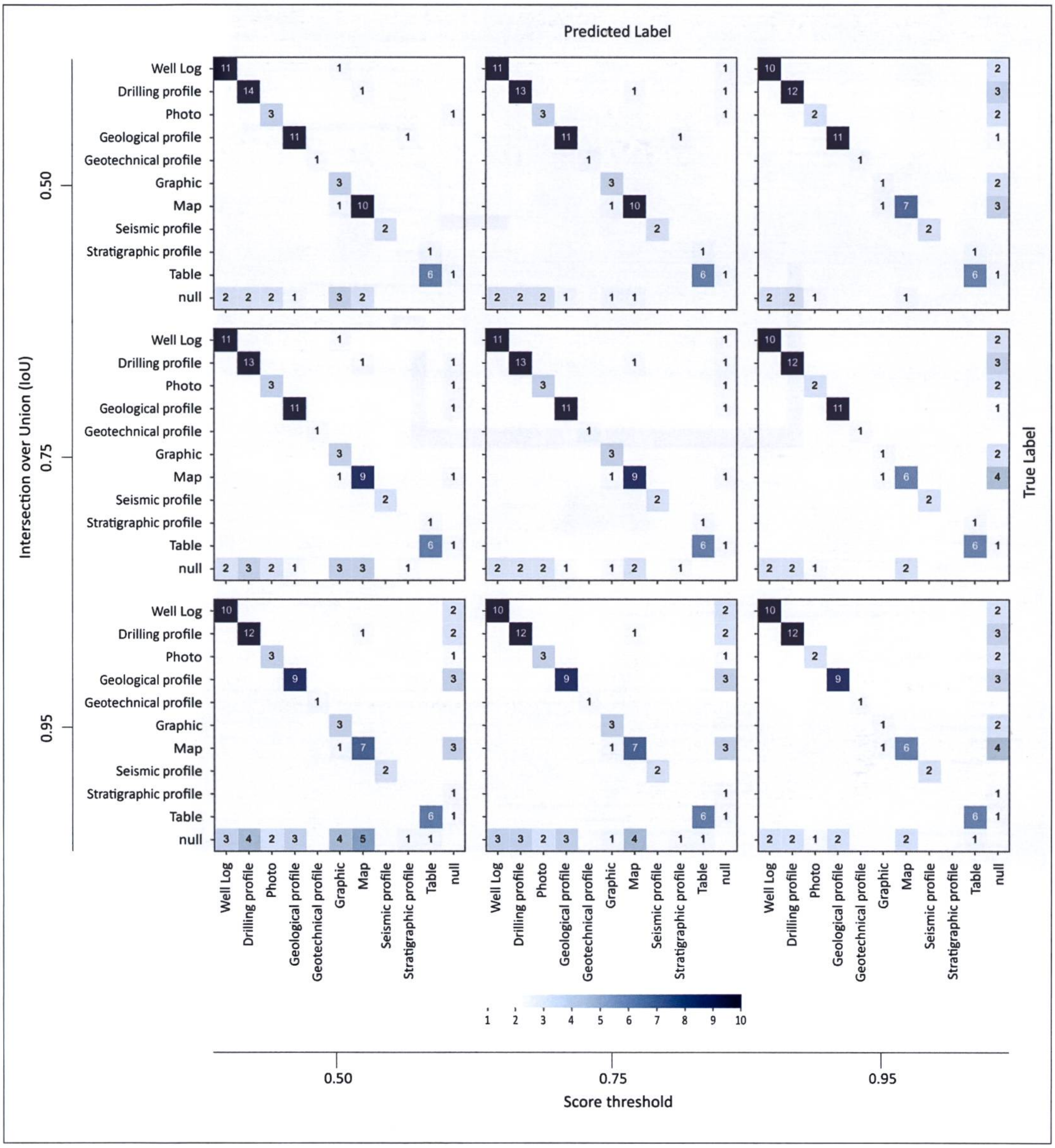

Fig. 3: Confusion matrices for Intersection over Union (IoU) range between 0.5 to 0.95 and score threshold range from 0.5 to 0.95 for the object detection task.

contain legends and ground truth bounding boxes might be difficult to annotate. On the other hand, all other classes show confident APs of greater than 70.

By taking the Intersection over Union (IoU) into account, AP with IoU > 0.5 (AP50) and IoU > 0.75 (AP75) are calculated (Tab. 4). The IoU indicates how accurate the model is compared to ground truth object markings. Thereby, it appears that AP50 (77.39) achieves only a marginally better AP than AP75 (73.96), indicating that the model accurately marks detected objects.

A similar pattern is observed, when looking at the confusion matrix with an IoU range from 0.5 to 0.95 and score thresholds from

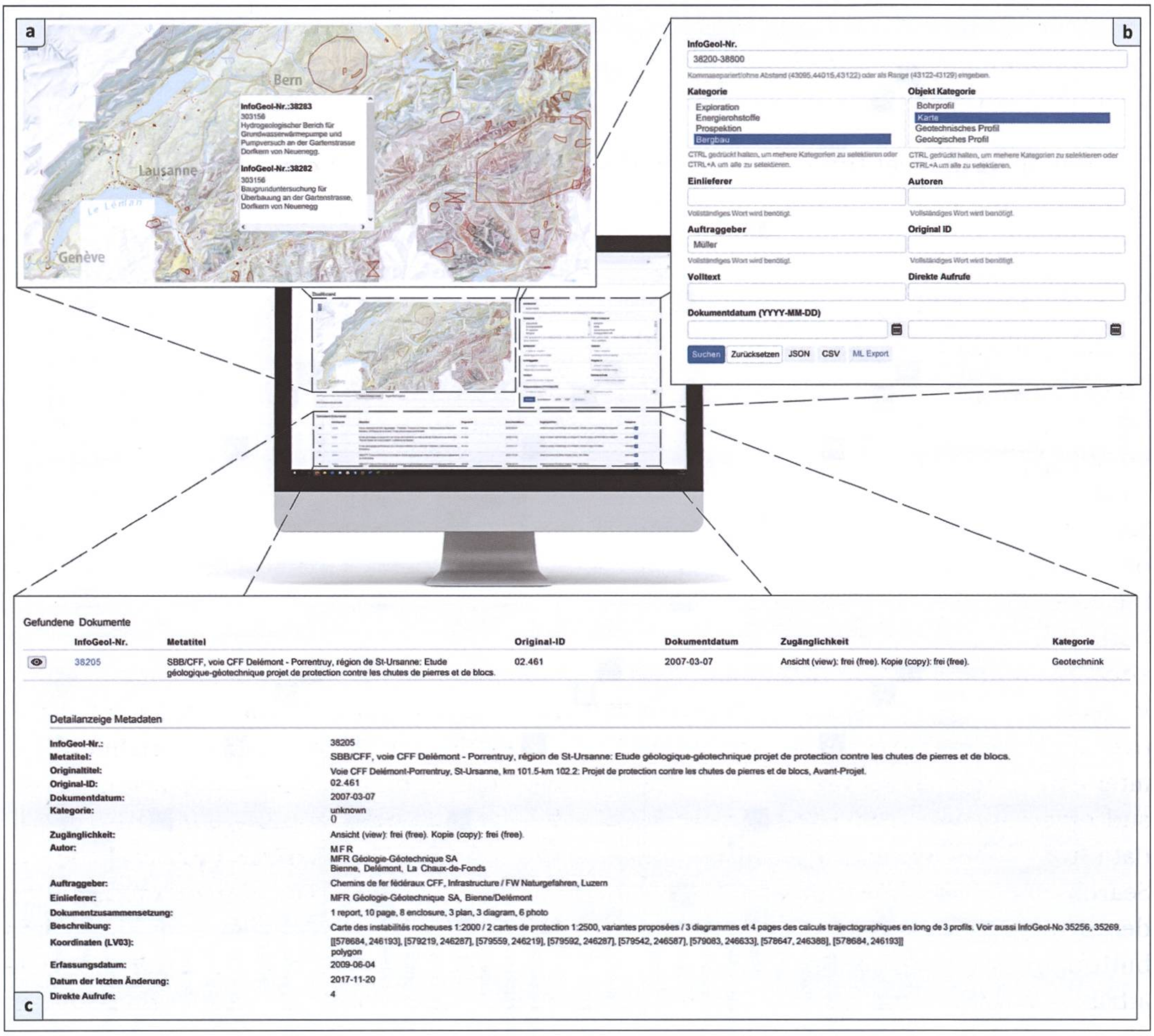

Fig. 4: Engineered web application for querying meta data specifically designed for the needs of the SGS (a) map viewer including area filter functionalities and geometry tool tip, (b) meta data filter specifically designed to the users need including export to csv and json and (c) document list section corresponding to filters from (a) and (b) with «eye» button for advanced information as well as further link to pdf files and download options.

0.5 to 0.95 (Fig. 3). The score threshold indicates how accurate the model is with respect to true label. The first confusion matrices column on Figure 3 shows increasing IoU with constant score threshold. In this case, the model prediction is already fairly accurate with low IoU. With increasing IoU, i.e., with increasing restriction, more elements are assigned to the null category as it is expected since it is easier to predict objects with low IoU. Most prediction errors occur due to the accuracy of the bounding box and not due to wrong classification of the detected object. On the other hand, we can emphasise that the model predicts remarkably many false-positive objects in the null category (true label) almost independent of IoU, score threshold or specific object classes. Moving towards higher score thresholds, the overall pattern stays the same. However, the remaining false-positive and false-negative predictions tend to move towards the null class, almost independent of the IoU. This shift can be explained due to the increasing restriction of both parameters, IoU and score threshold.

Web Application

The web application is basically developed in order to be able to visualise and query the results of the above described pipelines. It is designed to meet the needs of the SGS. Existing meta data of the current meta data management application (Hayoz et al. 2009) are used together with our newly created (meta) data from the two machine learning models. Basically, the application dashboard consists of three major elements, displayed in Figure 4: a) Geographical map view, b) attribut filter and c) table view of search results. For querring data, a geographical map search for filtering geological assets based on its loacation (Fig. 4a) as well as an attribute filter (Fig. 4b) is implemented. Of course, both search masks can be combined as desired. In addition, further details on a specific geometry may be displayed by hovering over individual geometries in map view. The filters with drop down menu for document and object classes refer to newly generated data from the two machine learning models. Search results are shown in a table view, ordered by ID (Fig. 4c). By clicking the «eye» button, detailed information comprising all attributes of the respective asset can be explored. Moreover, it is possible to directly view and explore each page and querying for specific objects. Finally, there are also download functions, on the one hand for meta data as json or csv and, on the other hand, a direct download of the corresponding geological asset as pdf.

4 Conclusions

We engineered key tools for processing scanned geological archive data. Our two processing pipelines can take on three important tasks on the digitisation path towards fully digitised and structured subsurface data. The OCR process generates digital base data after scanning the analogue data and paves the way for further digital processing and reuse of previously unstructured data. The introduced deep learning algorithms build on this foundation and offer the possibility of carrying out detection and classification tasks towards structured geological data. Automatised categorisation to specific geoscientific classes offer the possibility to query only relevant topics and methods for answering a desired geological question. We think that with simultaneous searches for specific objects, such as stratigraphic profiles or maps, the gathered data open up new and easy searchability in the field of geology. The geological archive of the SGS can be significantly revaluated with the combination of the engineered pipelines, web application and the already existing meta information for geological assets.

However, the proposed methods have limitations and room for improvements. The most relevant model performance limitation can mainly be attributed to the small number of labelled data per geoscientific class, both for the text classification and the object detection task. For both tasks, this number is clearly at the lower end. More data will most likely improve the results. However, with the actual data volume, it is difficult to predict a scaling effect when adding more labelled data. Nevertheless, we assume that it would be possible to improve the results by doubling the labelled data. At the same time, it would be possible to assess the extent to which a further increase in data has an impact on the model performance. Moreover, the text classification is limited by the rather ambiguous definition of the classes. Other experts (i.e., geoscientists) will perhaps not agree with the classifications. This fact can also lead to considerable difficulties when another expert labels new data. There may be inconsistencies in the training set and the model becomes worse instead of better. This indicates that unique classes with little room for interpretation are important. In any case, even with measures like this, it is impossible that all experts completely concur with the classifier and the classification. This clearly highlights the heterogeneity in geological data as well as how and by whom geological data is interpreted. Nevertheless, this current release based on small training datasets shows already excellent results benchmarking today's capabilities.

5 Outlook and Further Applications

Our study represents the initialisation of the digital transformation with the employment of artificial intelligence at the Swiss Geological Survey (SGS). During the next development steps, our proposed methods will be operationalised, which shall lead to a next generation data management process. The idea is to facilitate the digitisation of geological archives and by this to unlock their extensive and valuable information. For that, the SGS plans to incorporate the proposed methods into online digitalisation tools, which may support the digitisation process of other geological archives, for example from federal offices, cantons, or private companies. In fact, the aim is to provide tools for processing unstructured scanned geological data and to support the geoscience community on the digitalisation pathway without any direct data gathering benefit for the SGS. The final goal is to provide the public easy access to the possibility to digitise geological data. This will lead to a better understanding of the geological subsurface as access to digital and analysed data will be much easier than to analogue archives. In addition, the proposed pipelines are fully open-source and can be used to process any geoscientific archive and transfer the scanned (analogue) rudimentary managed digital document repositories into a fast, simple and performant search engine.

Author's Contribution

JMo: project lead, conceptualisation, data processing, visualisations, writing, coordination. DFr: development OCR, text classification and object detection pipeline. JMe: development and deployment web application. NOe: conceptualisation, coordination data labelling at the SGS, review. GPe: project management, conceptualisation of pipelines, visualisations, writing, review. SHe: review. All authors read and approved the final manuscript.

References

Bergstra, J. & Bengio, Y. 2012: Random search for hyper-parameter optimization: Journal of Machine Learning Research, v. 13.

Brodhag, S. 2022: Geologisches Informations- und Produktionssystem GIPS – Schlussbericht TP5 Geologische Assets: Unpublished report, Swiss Federal Office of Topography swisstopo (in preparation).

Davies, E. R. 2012: Computer and Machine Vision, Computer and Machine Vision.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. 2019: BERT: Pre-training of deep bidirectional transformers for language understanding, NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference, Volume 1.

Django Software Foundation 2013: django - The web framework for perfectionists with deadlines: https://djangoproject.com, Accessed 01 February 2022.

FARM 2021: FARM: Framework for Adapting Representation Models. https://farm.deepset.ai/, Accessed 01 February 2022.

Gormley, C. & Tong, Z. 2015: Elasticsearch: The Definitive Guide.

Hayoz, P., Pfiffner, M., Schaller, T. & Schwyter, B. 2009: InfoGeol - Software Architecture Document: Unpublished report, Swiss Federal Office of Topography swisstopo.

Kay, A. 2007: Tesseract: An Open-Source Optical Character Recognition Engine: Linux J., v. 2007, p. 2, Belltown Media.

Merkel, D. 2014: Docker: lightweight linux containers for consistent development and deployment: Linux journal, v. 2014, p. 2.

Oesterling, N., Brodhag, S. & Morgenthaler, J. 2022: Datenmodell Geologische Assets: Swiss Federal Office of Topography swisstopo. https://github.com/swissgeol/dmAssets, Accessed 01 February 2022 (in preparation).

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J. & Soumith, C. 2019: PyTorch: An Imperative Style, High-Performance Deep Learning Library: Advances in Neural Information Processing Systems 32, p. 8024-8035, Curran Associates, Inc.

Swiss Council 2018: Geologische Daten im Untergrund. Bericht vom 07.12.2018, Bern, p. 24 https://www.are.admin.ch/are/de/home/medien-und-publikationen/publikationen/strategie-und-planung/geologische-daten-zum-untergrund.html, Accessed 01 February 2022.

Sekachev, B., Zhavoronkov, A. & Manovich, N. 2019: Computer Vision Annotation Tool: a universal approach to data annotation: Intel.

swisstopo 2021: Digitalisierung des geologischen Untergrunds - Umfeld, Herausforderungen und Massnahmen. Bericht an den Bundesrat, p. 38, https://www.admin.ch/gov/de/start/dokumentation/medienmitteilungen.msg-id-83491.html, Accessed 01 February 2022.

Wu, Y., Kirillov, A., Massa, F., Lo, W.-Y., Girshick, R., Wu, U., Kirillov, A., Massa, A., YenLo, W. & Girshick, R. 2019: Detectron2: https://github.com/facebookresearch/detectron2, Accessed 01 February 2022.