Distributed Skin Lesion Analysis Across Decentralised Data Sources

Yongli MOUa*, Sascha WELTENa*, Mehrshad JABERANSARYa, Yeliz UCERYEDIELb, Toralf KIRSTENC, Stefan DECKERa,b and Oya BEYANa,b,1

a RWTH Aachen University, Germany
b Fraunhofer Institute for Applied Information Technology (FIT), Germany
c Hochschule Mittweida, Germany

Abstract. Skin cancer has become the most common cancer type. Research has applied image processing and analysis tools to support and improve the diagnose process. Conventional procedures usually centralise data from various data sources to a single location and execute the analysis tasks on central servers. However, centralisation of medical data does not often comply with local data protection regulations due to its sensitive nature and the loss of sovereignty if data providers allow unlimited access to the data. The Personal Health Train (PHT) is a Distributed Analytics (DA) infrastructure bringing the algorithms to the data instead of vice versa. By following this paradigm shift, it proposes a solution for persistent privacy-related challenges. In this work, we present a feasibility study, which demonstrates the capability of the PHT to perform statistical analyses and Machine Learning on skin lesion data distributed among three Germany-wide data providers.

Keywords. Distributed analytics, personal health train, federated learning, image processing

1. Introduction

In recent years, melanomas have emerged as the most common cancer type. Rising incidences sensitise the public to skin cancer, which has become a growing public health concern [1,2]. At this point, dermoscopy has become a tremendously successful diagnostic tool to support dermatologists avoiding a harmful or even lethal disease progression in an early stage [3]. However, clinicians need years of expertise to classify possibly malicious skin lesions correctly, which hampers the efficiency and effectiveness of such a treatment. Therefore, recent research focused on decision support systems encompassing model-based approaches for automated lesion detection.

The conventional development of Machine Learning (ML) models requires access to vast amounts of data, which usually must be centralised beforehand. This can pose two main challenges. First, the process of data centralisation potentially acts as a bottleneck complicating the development process. Second, centralising medical data does not often comply with new data protection regulations. The European GDPR introduces strict restrictions on the processing of personally identifiable data. Personal data can be collected, stored, and analysed according to the legal bases defined in Article

1 Corresponding author: E-mail: beyan@dbis.rwth-aachen.de
* Authors contributed equally.
6 of the GDPR [4]. Although GDPR does not limit the use of personal data for analytics, large-scale data collection entails the single point of failure and offers limited transparency and provenance [5]. This will lead to the loss of data sovereignty as well as the difficulty in complying with the GDPR. Methods to circumvent these challenges are decentralised approaches, which bring the analysis to the data instead of vice versa. This paradigm shift enables data analyses with data remaining inside institutional borders. Due to this circumstance, Distributed Analytics (DA) architectures propose solutions for preserving data privacy and complying with the above-mentioned existing regulations.

In this work, we present a feasibility study, which demonstrates the capabilities of DA on decentralised medical records. Precisely, we aim to perform statistical analyses and to develop an ML-based detection model on skin lesion patient data nationwidely distributed across three data providers. We find that DA constitutes indeed a valuable alternative to centralised approaches by simultaneously preserving patients' data privacy.

2. Methods

2.1. Personal Health Train

We resort to a DA infrastructure called Personal Health Train (PHT) as it has shown its promising capabilities of performing complex model training procedures in studies on decentralised lung cancer data or radiomics data [6-8]. The PHT supports data management, data analysis, and the responsible usage of sensitive medical records [6-10]. The key elements in the PHT ecosystem are the so-called Trains and Stations. The Train encapsulates the analysis tasks (algorithms, queries or models) by using containerisation technologies. The Train contains all prerequisites to query the data, execute the algorithm, and store the results. The Stations act as data providers that maintain repositories of sensitive health data. In order to analyse decentralised data, a Train is sequentially transmitted to every Station, which executes the Train. The Train performs the analytical task and calculates the results (e.g. statistics or model parameters) based on the locally available data. By visiting each Station in succession, the Train incrementally updates the analysis results until the final Station has been visited. An overview of the described setting is given in Figure 1.

2.2. Data Management

2.2.1. Dataset and Distribution

Our experiments are conducted on the skin lesion image dataset for melanoma classification from the International Skin Imaging Collaboration (ISIC) 2019 Challenge2. The used dataset consists of two data types. First, it includes 25,331 dermoscopic images. Each image is classified into one of eight different diagnostic categories indicating the lesion types. Second, each image is (partially incompletely) described with patient metadata, e.g., the approx. age and the anatomical site. In our scenario, we simulate the decentralised dataset among our three institutions as follows. First, we create a test set, which contains 5066 images covering 20% of the initial dataset. The remaining 8% are randomly split into three subsets, i.e., each Station obtains 6755 images.

---

2 The ISIC archive maintains a list of related works: https://www.isic-archive.com
2.2.2. Data Storage

Each Station provides one storage system for each data type. We use the Blaze FHIR server\(^3\) for the patient metadata management. Further, the clinical images are kept in a standalone object storage and management system. We use MinIO\(^3\) as such a file storage system providing the desired functionality of storing images at predefined locations. We run a Docker 19.03.8 Engine\(^3\) on every Station to deploy all data provision components as containers.

2.2.3. Data Model

Before we store the clinical metadata on the FHIR server, we transform the metadata into FHIR resources\(^3\). We model the ISIC Challenge as a so-called FHIR ImageStudy resource to give information under which circumstances the image has been created. Each patient described in the metadata is saved in a FHIR Patient resource capturing patient-related information. Each Patient resource is linked to a FHIR Media resource. This Media resource captures a URL referencing to the location of the patient’s dermoscopy image, which is stored at a different location on the MinIO server. This setup has the advantage that the FHIR server is the only interface for the Train to gather both images and metadata using one single query.

2.2.4. Data Access

Based on this preliminary work, the Train encapsulating the analytical task is accessing the data as follows (see Figure 1). During the container instantiation, the Station admin passes the FHIR connection credentials to the Train container. After the Train execution, it queries the Patient resources of the ImageStudy to obtain the URLs stored in the Media resource. The analytical task downloads the images from the MinIO server based on the given URLs and performs the actual data analysis. Based on the present data types, we can perform different analyses, which we will discuss in the upcoming section.

![Figure 1. Data Management and Access at each Station.](image)

2.3. Analytical Tasks

We split our study into 2 applications\(^3\) introducing different levels of complexity.

2.3.1. Statistical Analysis

We will perform a statistical analysis on the clinical data. The corresponding Train will first conduct a quality check in terms of attribute completeness at each Station.

---

\(^3\) Sample images, metadata, results, and code available at: https://github.com/rwth-i5/mie2021
Afterwards, the Train will explore the class distributions. Results are aggregated with those obtained from previous Stations and transmitted to the next Station.

2.3.2. Image Analysis

We develop a ResNet-18 model based on the PyTorch framework and pretrained on the ImageNet dataset for the classification of the skin lesion images [11]. The learning rate is set to $10^{-4}$ and we use the Adam optimiser with weight decay of $5 \times 10^{-4}$. At each Station, the model will be trained for 40 epochs on 80% (training set) and internally validated by 20% (validation set) of the available images. The final model is evaluated on the test set. Each task has been implemented in Python 3.6. For the execution at each Station, the analyses have been Dockerised for being compatible with the Station's Docker Engine.

3. Results

3.1. Statistical Analysis

Through the quality check, we found that 89% of the provided data instances at each station capture all attributes of approximate age, anatomical site and sex, respectively. Further, a distribution analysis has revealed that more than half of the data is classified as melanocytic nevus (NV). Melanoma (MEL), basal cell carcinoma (BCC), and benign keratosis (BKL) cover about 17%, 13%, and 10% respectively. The remaining four classes, i.e., actinic keratosis (AK), dermatofibroma (DF), vascular lesion (VASC), and squamous cell carcinoma (SCC) occur less than 10%. Moreover, the diagnostic categories are independent and identically distributed (i.i.d.) among the Stations.

3.2. Image Analysis

After the distributed model training, we evaluate the model according to our holdout test set. We calculate accuracy and recall to measure its quality. For comparison, we analogously train a model on the centralised dataset. The result shows that our PHT approach has a comparable accuracy to the centralised approach but a difference in recall is evident. The centrally trained model has 75.40% of accuracy and 69.22% of recall, which represents our upper bound. The final model we obtained from our PHT approach has achieved 71.83% of accuracy and 63.35% of recall.

4. Discussion and Conclusion

In this work, we have presented our methodology to conduct DA using the PHT infrastructure developed by the institutions RWTH Aachen, FIT Fraunhofer, and HS Mittweida. Exemplary statistical and image analysis use cases on artificially distributed skin lesion data across three Germany-wide geographically-separated Stations have been presented. Our study has shown the DA capabilities of performing data analysis with different complexity levels. Further, we have demonstrated that the quality of the decentralised model training is similar to the central model training. However, the latter relies on data centralisation, which does not comply with data protection guidelines and
therefore not applicable for many analytical applications. Our approach is an enabler for analyses on sensitive data by accessing data at its origin. Since only calculated results (e.g. aggregated statistics or model parameters) leave the institutional borders, the PHT constitutes a valuable option for privacy-preserving data analysis while simultaneously producing results, which are comparable to the centralised approach.

Nevertheless, DA poses some new and inherent challenges. Since the datasets are decentrally generated, the underlying distribution is not necessarily i.i.d. among the Stations limiting the effectivity and efficiency of the decentralised model training. One consequence of this circumstance is a phenomenon called Catastrophic Forgetting, which we also experienced in our experiments. We observed that the training loss increases, when the model starts the training at another Station. This means that the model periodically forgets already learned features, which causes a prolonged training duration. Furthermore, the present Station heterogeneity has emerged as a technical bottleneck since different amounts of computational resources have been provided.

As future work, we intend to apply the PHT infrastructure to more datasets and aim to reveals insights about the influence of highly-skewed distributions on the analysis outcome. To alleviate the influence of non-i.i.d. data, we strive to make conventional ML algorithms more robust and suitable for DA applications, e.g., by introducing additional layers for data curation. Further effort is put into the improvement of the transparency of the result calculation and the interoperability between several implementations to facilitate the access to more data.

References