What Your Radiologist Might be Missing: Using Machine Learning to Identify Mislabeled Instances of X-ray Images

Tim Rädsch
Karlsruhe Institute of Technology, Germany
timraedsch.research@gmail.com

Sven Eckhardt
Karlsruhe Institute of Technology, Germany
sveneckhardt.research@gmail.com

Florian Leiser
Karlsruhe Institute of Technology, Germany
florianleiser.research@gmail.com

Konstantin D. Pandl
Karlsruhe Institute of Technology, Germany
konstantin.pandl@kit.edu

Scott Thiebes
Karlsruhe Institute of Technology, Germany
scott.thiebes@kit.edu

Ali Sunyaev
Karlsruhe Institute of Technology, Germany
sunyaev@kit.edu

Abstract

Label quality is an important and common problem in contemporary supervised machine learning research. Mislabeled instances in a data set might not only impact the performance of machine learning models negatively but also make it more difficult to explain, and thus trust, the predictions of those models. While extant research has especially focused on the ex-ante improvement of label quality by proposing improvements to the labeling process, more recent research has started to investigate the use of machine learning-based approaches to identify mislabeled instances in training data sets automatically. In this study, we propose a two-staged pipeline for the automatic detection of potentially mislabeled instances in a large medical data set. Our results show that our pipeline successfully detects mislabeled instances, helping us to identify 7.4% of mislabeled instances of Cardiomegaly in the data set. With our research, we contribute to ongoing efforts regarding data quality in machine learning.

1. Introduction

Machine learning (ML) models can increasingly match or outperform humans at various tasks in diverse application domains. Contemporary ML models have, for example, been shown to identify lung diseases on chest X-rays more accurately than radiologists [1], or beat experienced players in notoriously difficult-to-master games [2]. Next to algorithmic advances, another fundamental reason for the recent progress in ML is the growing availability of training data [3], which is driven by the accelerating digitization of nearly all aspects of our everyday lives. This increasing availability of training data, however, is often contrasted by quality deficiencies of the data sets [4]. Especially when training data has to be labeled for supervised ML tasks, label quality is a common problem (i.e., instances are often mislabeled), as training data are mostly still labeled by hand [5]. Not only is such manual labeling of large data sets costly and time-consuming, but also error-prone [6].

Eventually, poor-quality training data might not only affect ML model performance negatively [6, 7], but also make post-hoc explainability and, therefore, trust in such ML models' predictions more difficult to attain, and may even inhibit the adoption of ML in information systems [8]. The problem of mislabeled instances in training data is further aggravated where expert knowledge is required to label data. Particularly in the medical field, one of the most promising domains for the application of ML [9], poor label quality is an immense problem [10]. On the one hand, medical experts are scarce, expensive, and their limited time is better spent on other tasks (e.g., treating patients) than on identifying mislabeled instances or labeling instances, to begin with [10]. On the other hand, the consequences of poorly labeled data and, thus, poorly performing ML models can be especially severe. They might negatively impact patients' health, such as misclassification of disease [11], enforce biases [12], or disadvantage the poor [13].

Extant research has dealt extensively with label quality in the context of supervised ML. In the past, researchers have especially focused on the ex-ante improvement of label quality by proposing improvements to the labeling process (cf. Section 2.1). More recently, researchers have also begun investigating means for the ex-post improvement of label quality, such as training ML models on the respective data sets and classifying instances in the data set with this trained model in order to identify mislabeled instances in training data sets automatically [5]. While such approaches generalize well across different data sets [5], they are usually inconsiderate of the specific characteristics of different data sets (e.g., the degree of confidence for an individual instance's label, or the

temporal dimension of a data set in which multiple instances exist for one patient). We, therefore, ask the following research question:

RQ: How can we improve label quality in medical data sets ex-post by means of using ML to detect potentially mislabeled instances?

To answer our research question, we build on recent research for the ML-based detection of mislabeled instances and propose a two-staged pipeline for the identification of mislabeled instances that, in addition, takes advantage of salient characteristics of the underlying data set. Specifically, we focus on the CheXpert data set (see Section 2.2 for a detailed description of the data set) and make use of—what we call—explicitly and implicitly labeled data instances, as well as the fact that the data set contains a temporal dimension (i.e., providing multiple instances for the same patient at different points in time) to better detect potentially mislabeled instances. In doing so, we provide a two-fold contribution. First, we propose a novel method that combines an ML approach in form of a convolutional neural network with a data set-specific heuristic to improve the detection of potentially mislabeled instances. Applying our method can improve the explainability and, thus, the trustworthiness of ML-based information systems [14]. Second, we directly contribute to the CheXpert data set by identifying potentially mislabeled instances of Cardiomegaly in the data set.

The remainder of this paper is organized as follows. In the next section, we briefly summarize the related research on improving label quality in supervised ML and describe the CheXpert data set that we use for our analysis. Afterward, we summarize our proposed, two-staged pipeline in section three, before we describe the first stage in more detail in section four, and the second stage in more detail in section five. We discuss our results in section six and conclude this paper with section seven.

2. Background

2.1. Data labeling and label quality

When data instances are acquired (e.g., images using a camera), they are usually unlabeled. In order for acquired data instances to be usable for supervised ML, they must be annotated first (i.e., one or several labels must be assigned to each instance). Although there exist approaches for the automatic extraction of labels (e.g., through rule-based natural language processing), most annotation tasks are either completely or at least partially performed by humans today [21]. Simple annotation tasks, for example, are often crowdsourced through platforms like Amazon's Mechanical Turk [22], whereas more complex annotation tasks are outsourced to dedicated teams in low-income countries [21, 23]. Yet, the manual annotation of large data sets is costly, time-consuming, and error-prone [6]. Typical sources of error in the annotation process include subjectivity, data-entry error, or inadequate information [6].

Extant research has dealt extensively with label quality in the context of supervised ML. Toward this end, researchers have especially focused on the ex-ante improvement of label quality by proposing improvements to the annotation process (sectors A and B in Table 1). Such approaches typically generalize well across different types of data sets for different industries, or are developed specifically for one type of data set in a specific industry. An example for a method that generalizes well across different types of data sets and industries is an annotator rating system that aims to maximize the label quality while minimizing the number of annotators required [15, 16, 17]. An example of a method specifically developed for one industry is the usage of citizen scientists to annotate bird images [24]. In the health care industry, approaches deal with a wide range of specific issues such as training data [18], image classification tasks [19] or reference data generation [20]. The majority of these approaches use crowdsourcing and a recent overview of the usage of crowdsourcing in medical image analysis is provided by Ørting et al. [25].

Alternatively, researchers have started to investigate ML-based approaches for the ex-post identification of mislabeled instances after these instances have been (incorrectly) annotated already (sectors C and D in Table 1) [5, 6, 7]. Again, the majority of contemporary approaches generalize well across different data sets for different industries, for example, by training an ML model on the labeled data set and reclassifying the instances of the same data set, which uncovers potential label errors [5]. In contrast, we develop our approach specifically for the peculiarities of the CheXpert data set as a representative in the medical imaging industry, which especially includes the presence of explicit and implicit labels, as well as a temporal dimension in the data set as described in Section 2.2.

2.2. The CheXpert data set

Medical imaging is one of the major application domains for supervised ML [10], with chest X-rays being the most common medical image diagnostic worldwide today [26]. Medical imaging is also a domain in which mislabeled instances can have particularly

severe consequences. Accordingly, we decided to focus on the medical imaging domain in this paper and draw on the CheXpert data set to develop a system pipeline for the identification of potentially mislabeled instances in large ML data sets [27]. In the following, we provide a brief description of the data set.

The CheXpert data set was compiled by a Stanford-based research group led by Jeremy Irvin and Pranav Rajpurkar. It contains 224,316 instances of chest X-rays, 85% of which are frontal scans, while the remaining 15% are lateral scans. Each instance in the data set has a variety of corresponding labels, including a patient ID (total: 65,240 patients), patient age and sex, and 14 labels that represent clinical observations (i.e., no finding, support devices such as pacemakers, and 12 medical conditions such as fractures, pleural effusion or Cardiomegaly) of a given X-ray scan. For each instance in the data set, each of these 14 labels contains either a 1 (i.e., the clinical observation corresponding to that label is present in that scan), a 0 (i.e., the clinical observation corresponding to that label is not present in that scan), a -1 (i.e., there is no clear indication of the presence or absence of the corresponding disease in that scan), or NaN (i.e., there is no information available about the corresponding clinical observation for that scan). Further on, to provide a mapping between the scans and the corresponding labels, a *path* label, which assigns a unique scan to each instance in the data set, is included.

It is noteworthy that labels for the 14 clinical observations were constructed using rule-based natural language processing algorithms, which extracted relevant information out of radiologists' reports. NaN, therefore, has a dual meaning. Either the radiologists did not investigate this observation due to a clear lack of its presence, or the rule-based natural language processing algorithms were unable to identify or translate the aspect in the respective report. In both cases, a thorough investigation of label correctness should result in an improved label quality.

Several studies have made use of the CheXpert data set. This includes, for example, studies testing different approaches toward explainable artificial intelligence in radiology [28], studies aiming to detect multiple diseases on a single X-ray scan [29], or studies analyzing the computational efficiency and accuracy of different neural network architectures for radiology [30]. Furthermore, recent studies create COVID-19 disease detectors on chest X-ray scans and use the CheXpert data set to obtain negative instances [31] or for pre-training [32].

3. Pipeline overview

In order to detect potentially mislabeled instances in the CheXpert data set, we propose a two-stage pipeline (see Figure 1). Within the first stage, and similar to Müller and Markert [5], we develop an ML approach, where we train a neural network on a high-quality subset of the main data set (i.e., the CheXpert data set) and then use the resulting ML model to reclassify instances in a test subset. Within the second stage, we develop a heuristic approach that builds on the results of the reclassification and applies additional, data set-specific criteria to identify potentially mislabeled instances more accurately. The ultimate objective of the heuristic approach is to further minimize error rates when identifying mislabeled instances.

Table 1: Overview of extant research streams.

		Approach	
		Improve annotation process (ex-ante label quality improvement)	**Identify mislabeled instances (ex-post label quality improvement)**
Industry	**General**	**[A]** Typical RQ: How to efficiently use human annotators? Example study: [15, 16, 17]	**[C]** Typical RQ: How to identify mislabelled instances in a data set? Example study: [5, 6, 7]
	Specific	**[B]** Typical RQ: How can multiple annotations be applied to assure highly reliable reference correspondences for endoscopic image? Example study: [18, 19, 20]	**[D]** RQ: How can we improve label quality in medical data sets ex-post by means of using ML to detect potentially mislabeled instances? *This study.*

4. Machine learning approach

The first stage of our pipeline consists of an ML-based approach that we illustrate in Figure 2. The grey highlighted areas represent all steps where we handle explicit data whilst white cells use implicit and explicit data as well. Within the approach, we train an ML model on a high-quality subset of the main data set and then use the resulting model to re-classify our testing data. If the re-classification result deviates from the instance's original label, we consider it as an indication for a mislabeled instance and further process this instance in stage two. In the following, we describe our ML approach and its outputs in more detail.

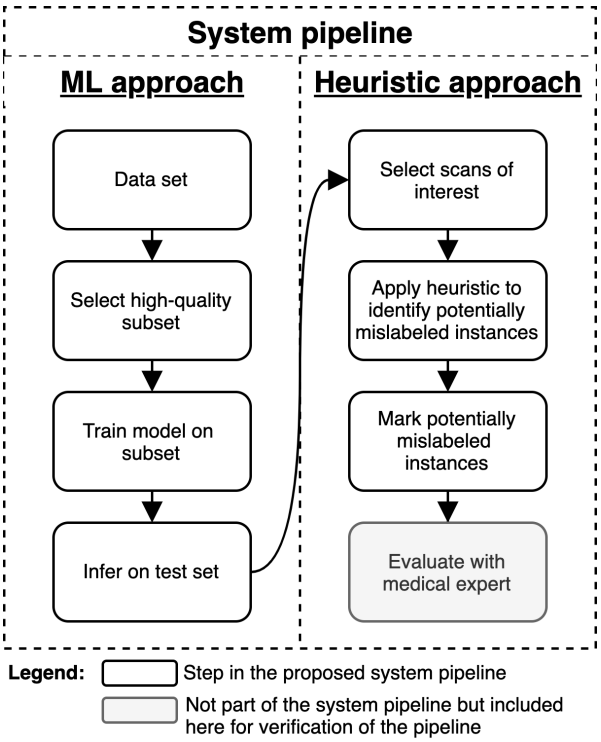


Figure 1: Overview of the proposed two-staged pipeline.

4.1. Data pre-processing

To decrease the likelihood that our ML model makes an error, we need to train it on data instances with high-quality labels. Accordingly, we introduce the distinction between explicitly and implicitly labeled instances. For a certain condition, each data instance is either explicitly labeled or implicitly labeled. For an instance to be explicitly labeled for a given condition, the condition needs to be mentioned in the radiologists' reports, assuming a value of -1 (no clear indication), 0 (condition not present), or 1 (condition present). In contrast, if there was no information on that specific condition in the radiologists' reports at all (i.e., the instance has a NaN label for that condition), it is considered implicitly labeled. Table 2 illustrates the distribution of implicitly and explicitly labeled instances in the CheXpert data set.

Out of the 14 different conditions in the CheXpert data set, we focus on the Cardiomegaly condition, which describes an enlargement of the heart. We do so for multiple reasons. First, Cardiomegaly has a high clinical importance with a prevalence of nearly 5.8 million people in the United States alone [33]. Second, it is also one of the five observations selected for the CheXpert competition tasks [27]. Third, among the competition tasks, Cardiomegaly has the highest percentage of implicit labels with 79.32%, thus, having great potential for improvement. Accordingly, the final instance labels for this study consist of *path*, *age*, *patient_ID* and *cardiomegaly*.

Before we can build and deploy our ML-based approach, we need to create different subsets out of our data set. For this, we group the X-ray scans on a per-patient basis. This ensures that each patient only occurs in one subset. We further filter out lateral scans, which leaves us with frontal scans, representing the majority of instances in the CheXpert data set (n = 191,027). We apply the common 70:30-split for machine learning algorithms, resulting in 133,718 frontal scans in the training data set and 57,309 frontal scans in the testing data set. We then modify the training

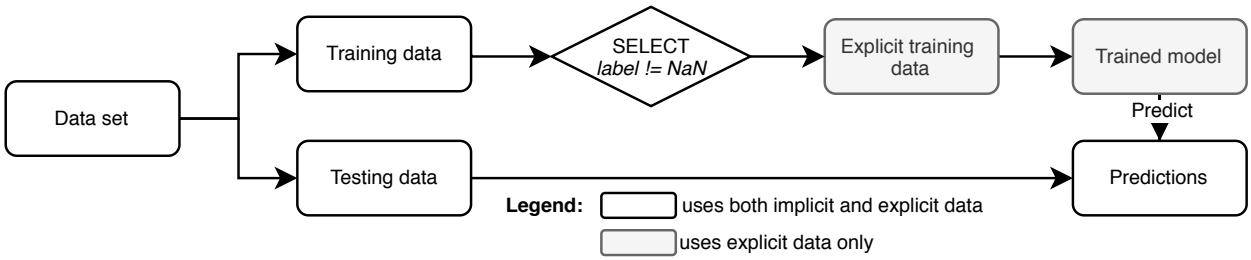


Figure 2: The workflow of the machine learning approach.

Table 2: Distribution of implicit and explicit labels in the CheXpert data set.

Medical condition	Implicit labels	Explicit labels	1.0	0	-1.0
No Finding	89.98%	10.02%	100.00%	0.00%	0.00%
Enlarged Card.	79.93%	20.07%	24.08%	48.26%	27.66%
Cardiomegaly	79.32%	20.68%	58.44%	24.06%	17.50%
Lung Opacity	47.28%	52.72%	89.64 %	5.60%	4.75%
Lung Lesion	94.65%	5.35%	76.91%	12.46%	10.63%
Edema	61.53%	38.47%	60.78%	24.11%	15.11%
Consolidation	68.39%	31.61%	20.93%	39.79%	39.28%
Pneunomia	87.64%	12.36%	21.87%	10.14%	67.99%
Atelectasis	69.36%	30.64%	48.76%	1.94%	49.30%
Pneumotorax	64.67%	35.33%	24.64%	71.38%	3.98%
Pleural Effusion	40.37%	59.63%	64.70%	26.57%	8.73%
Pleural Other	97.09%	2.91%	54.27%	4.87%	40.87%
Fracture	94.54%	5.46%	74.13%	20.60%	5.26%
Support Devices	44.85%	55.15%	94.14%	4.98%	0.88%

data set as follows.

We consider uncertain instances as negative instances because, given a positive prediction of Cardiomegaly from our model, we want to be as certain as possible that this prediction is correct and our model only learns explicit positive labels as such. In the case of a binary prediction, this handling resulted in the best performance for Irvin et al. [27]. To obtain the training set, we first filter it for explicitly labeled instances, leaving us with 26,407 explicit scans in the training subset. Of those scans, 61.64% were assigned positive labels, while 38.36% were assigned negative labels. Following the common rule of thumb, we then use 10% of the training data as a validation data set to implement our early stopping approach and prevent overfitting of our ML model. This results in 23,767 explicitly labeled scans for the training data set and 2,640 explicitly labeled scans for the validation data set. In doing so, our approach significantly differs from other studies using the CheXpert data set, which usually map the implicit data instances to a 0 label (negative finding) [27].

Additionally, we need a data set to generate metrics and create comparable results to the available literature, called the benchmark set. For this, we use the available CheXpert validation set. Together with the CheXpert testing set, which is not publicly available, it is annotated by five board-certified radiologists with a majority vote [27]. In total, we have 4 subsets of our original data set: the training set, the testing set, the validation set, and the benchmark set (not displayed in Figure 2).

4.2. Neural network architecture

After pre-processing our data, we now turn to the training of our ML model on the training data set. For this, we use the DenseNet-121 neural network architecture [34], which is widely used in the computer vision community. Additionally, prior research compared different architectures and discovered that the DenseNet-121 architecture works best for the CheXpert data set [27]. The DenseNet-121 consists of four *dense blocks*, which themselves consist of a convolutional layer followed by a pooling layer. The novelty introduced by the dense blocks are the direct connections from one layer to all following layers within the single dense block. This improves the information flow between the layers [34].

As we now have decided for a neural network architecture, we also need to set the hyperparameters, before we can train our neural network. For most hyperparameters, we orient ourselves toward prior research [27]. Specifically, we use the Adam optimizer with default β-parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$. Our learning rate is 1×10^{-4}. We keep the input image size of 320×320 pixels. However, as we have fewer images to train on, we decrease the batch size to 8. Furthermore, we implement an early stopping with the validation set to choose the right amount of epochs to train and prevent overfitting on the data.

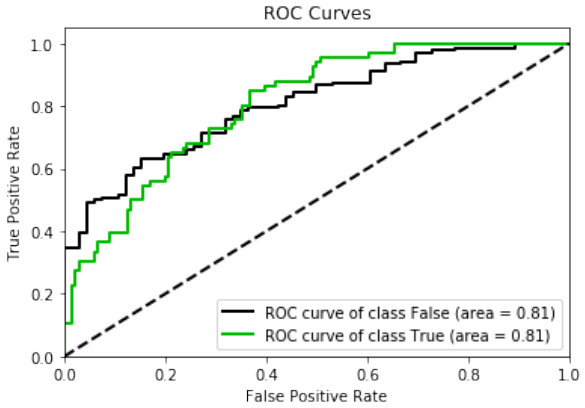


Figure 3: ROC curves of our trained ML model on the the benchmark set.

4.3. Evaluation of the neural network

In accordance with the implemented early stopping approach [35], we evaluate the performance of our network on the validation set after each training epoch. Since overfitting occurs if we train for more than 12 epochs, we stop our training after 12 epochs.

When evaluating the trained model with the benchmark data set, the model performs well with an area under the curve (AUC) of about 0.81 (see Figure 3), which is below the AUC of 0.9 reached in the original CheXpert study [27]. A potential reason for this could

be the fact that the benchmark data set is relatively small with only 200 scans and contains a different label distribution than the training data set. Still, we deem an AUC of 0.81 to be a value we can work well with, as our goal is not to improve the already good classification results of earlier studies but rather achieve a model with comparable results to be able to test our heuristics on.

5. Heuristic approach

5.1. Overview

After training the ML model on explicit instances only, we can employ the resulting ML model to classify implicitly labeled data instances on the testing set. If the classification with the ML model deviates from the instance's original label, either the instance may be mislabeled, or the model provided a wrong classification. To minimize the likelihood that the ML model provided us with a wrong classification, or in other words, to minimize the number of false positives, we developed a heuristic for the CheXpert data set that helps us with narrowing down the potentially mislabeled instances with confidence. We illustrate our heuristic approach in Figure 4 and describe it in more detail in the following.

5.2. Scans of interest

As we later in our heuristic want to identify patients who potentially have a Cardiomegaly diagnosis later in their life, we only focus on patients that have at least two points of diagnosis in our testing set. This leaves us with 25,655 scans containing both implicitly and explicitly labeled data instances in the testing set. The majority of patients removed from the testing set have only one scan provided in the entire data set.

Before we can classify scans of this testing set and detect potentially mislabeled scans, we need to transform the neural network output into discrete classes. By default, our model does not directly classify an image, but rather predicts a two-dimensional array whose values add up to 1. The first position of the array states the certainty of our model that an image has a positive finding for Cardiomegaly, whereas the second position states the certainty for a negative finding. To obtain discrete class labels, we need a threshold that separates positive and negative classes. A simple approach is to classify a class prediction as positive once the likelihood for it is predicted to be higher than the opposite likelihood. That means we would set the threshold to 0.5 and if the model is more than 50% certain that the image belongs to a class, we also predict it as said class. However, this might not always work in practice. Different types of errors incur varying opportunity costs in real-world scenarios [36]. Our approach identifies potentially mislabeled scans that should be checked by a radiologist and ideally be relabeled. Currently, one of the relevant issues for medical training data is the availability of medical personnel for labeling [10]. Choosing a lower threshold for predicting an implicit instance as positive would result in a higher recall but lower precision. Likewise, choosing a higher threshold would result in a higher precision, but it would also miss some mislabeled instance in the data set.

Consequently, we analyze possible thresholds on the benchmark set and observe that with a threshold value of *0.7864147* we have no false-positive predictions on the benchmark set. With this value, we are able to minimize the false positive predictions, while still having good results. After setting our threshold based on the benchmark set, we test our threshold on our testing data set. We use the explicitly labeled data instances to assess the performance of our model on the testing set. Overall, we have 8,600 explicitly labeled data instances in our testing set. On these explicitly labeled data instances, we obtain 180 false-positive predictions with our model, which is a rate of about 2%. Considering our goal of minimizing the false positive rate, we deem this to be a good result and, therefore, selected the threshold accordingly.

The remainder of the testing set consists of implicitly labeled data instances. These are interesting for our heuristics, and we use them to find potentially mislabeled data instances. In total, there are 17,055

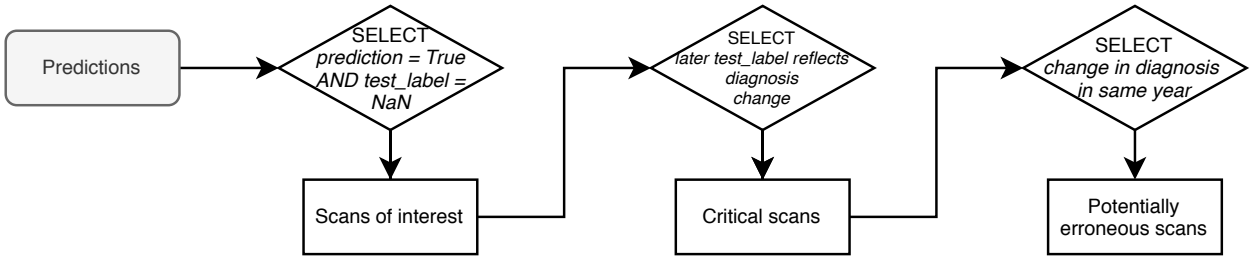


Figure 4: The heuristic approach we follow to identify potentially mislabled instances.

implicitly labeled data instances on the testing set. Out of these, our model predicted 1,500 instances (spread amongst 691 patients) to have a positive finding for Cardiomegaly while originally having a NaN label on the testing set. Therefore, these scans are potentially mislabeled and we further investigate these in our heuristic. We refer to these as *scans of interest*.

5.3. Critical scans

In order to maximize the precision of our approach, we further filter the acquired scans of interest and only focus on patients whose indication for Cardiomegaly changed between their examinations. In total, 662 patients had a change in their Cardiomegaly diagnosis, resulting in 1,431 potentially mislabeled scans. As these scans have an even higher likelihood to be mislabeled, we call these *critical scans*. Nearly all of the *scans of interest* are also *critical scans*.

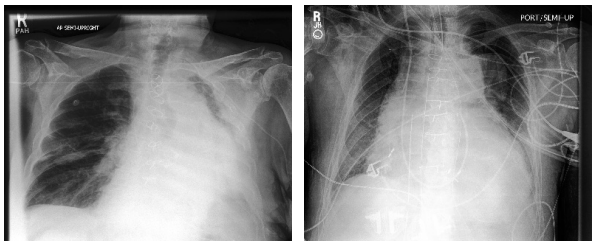
5.4. Potentially erroneous scans

To even further increase the precision of our heuristic, we now focus on scans where a change in diagnosis happened within one year and not during the lifetime of the corresponding patient. Since each scan comes with a label on the patient's age at the time of the scan, we can obtain the time difference between two scans for a given patient. We refer to the remaining scans as *potentially erroneous scans*. This further increases the likelihood that a scan is mislabeled since a radiologist explicitly labeled the patient's scan shortly after. This leads us to 1,265 *potentially erroneous scans* (among 613 patients).

5.5. Evaluation

To verify the results of our pipeline, we conducted a blind study with a professional radiologist. Specifically, we provided the radiologist with 50 X-ray scans that we asked him to label on a scale from -5 (condition definitely not present) to 5 (condition certainly present) for the medical conditions Atelectasis, Cardiomegaly, Pneumothorax, and Consolidation. Within these 50 scans, 30 were selected from the detected *potentially erroneous scans*. Thereby, the 10 scans with the highest probability were selected by hand and 20 scans were selected at random out of the remaining potentially erroneous scans. We picked the other 20 scans randomly from the other conditions to avoid a bias in the set.

Out of the 50 scans, 34% were assigned a label of +3 or higher for the presence of Cardiomegaly, 46% were assigned a label of -3 or lower for the presence of Cardiomegaly, the remaining ones have a label somewhere in between. In total, the average label for the presence of Cardiomegaly is -0.66. Just looking at the 30 potentially erroneous scans, 17 or 56.7% have a label of at least +3 and only 8 or 26.7% have a value of -3 or lower. Two of the likely mislabeled scans are shown in Figure 5. Based on the neural network output, these two had the highest probability of being mislabeled, and the consulted radiologist confirmed that it is very likely that the patients suffer from Cardiomegaly. The remaining values in this set of 30 potentially erroneous scans are between the two above numbers. The average label within this set is +1.59.



(a) Patient 02001 study 18, view 1 **(b) Patient 37546 study 1, view 1**

Figure 5: Potentially mislabeled scans as confirmed by a professional radiologist.

Furthermore, all but one of the scans with a radiologist's label of at least +3 have been from the detected potentially erroneous scans, while most labels connoted with -3 or lower are from the randomly sampled scans. We, therefore, conclude that the radiologist's feedback confirms the effectiveness and precision of our approach in successfully identifying mislabeled data instances for medical image purposes.

6. Discussion

6.1. Principal findings

In this research, we propose a two-staged pipeline for the identification of potentially mislabeled instances in a large medical data set. For this, we first introduce the distinction between explicitly and implicitly labeled instances, which commonly occur in medical environments. We train an ML model on a high-quality subset of explicitly labeled instances and apply this model to re-classify data instances for which there was a higher uncertainty regarding the correctness of their labels. Subsequently, in the second stage of our pipeline, we develop and apply a heuristic that considers certain characteristics underlying our data set to further improve the certainty with which we are able to identify mislabeled instances. We evaluate the performance of our proposed system pipeline by surveying a professional radiologist who manually rated the likelihood of certain diseases on a set of X-ray scans.

This evaluation confirms that our approach successfully identifies mislabeled instances in medical data sets with high precision.

In total, we found 1,265 *potentially erroneous scans* in 17,055 implicit data instances, indicating that about 7.4% of the implicit instances are potentially labeled incorrectly in the CheXpert data set. Our results, therefore, slightly exceed error rates of 3% to 5% that are reported in earlier studies on label quality in public ML data sets [5, 37]. However, those error rates are reported on data sets in a broader, general context [38, 39, 37]. Looking at extant research, a variety of possible reasons for the occurrence of mislabeled instances might exist for our particular case. First, performing medical diagnoses is a difficult and error-prone task [40, 41], and consequently, radiologists might simply have missed a condition during the examination of some X-ray scans [42]. Second, the pooling of scans from different hospitals and patients, as is the case with the CheXpert data set, might have also resulted in some mislabeled instances. Under regular circumstances, patients are checked for a certain condition only and not for the full range of labels the CheXpert data set provides. Third, some instances might have been mislabeled due to inaccuracies in the rule-based label extractor (i.e., the condition was present in the text-based radiology report but not recognized and transferred to the label file). Toward this end, considering that our error rate only applies to Cardiomegaly, which is one of several diseases included in the original CheXpert data set, as well as our careful selection process and threshold, we can assume that the total number of incorrectly labeled scans is even higher.

Different steps in our heuristic reduce the number of identified mislabeled scans from 1,500 scans of interest to 1,265 potentially erroneous scans. As such, after applying our heuristic to filter down the scans and increase the precision, about 88% of the originally found *scans of interest* are also *potentially erroneous scans*. This suggests that even by applying only part of our heuristic, for example by only obtaining scans of interest and skipping the subsequent filtering steps, our approach is capable of precisely identifying a fairly large number of mislabeled instances. One likely reason for this is the fact that we trained our ML model on high-quality, explicitly labeled instances only, and therefore our initially trained model already performs quite well. Nevertheless, it also shows that by augmenting the first stage of the presented pipeline with a heuristic that caters to the specifics of the underlying data set, we were able to effectively lower the number of instances to be considered for relabelling.

6.2. Implications for research and practice

For research, we highlight the effectiveness of ML-based approaches for the ex-post detection of mislabeled instances. In particular, we show that multi-stage pipelines, such as combining ML-based approaches for the detection of mislabeled instances with data set-specific heuristics, can be useful for detecting mislabeled instances more accurately and with more confidence. This is especially useful in contexts where expert knowledge is required to (re-)label an instance and where such experts' valuable time is better spent on other tasks (e.g., health care) [10]. We can present to these experts the data instances which are most likely to be mislabeled and, thus, a correction can have the highest influence on improving the data set and building trustworthy ML-based information systems with that new data set. Additionally, we introduce the notion of explicitly and implicitly labeled instances and emphasize the importance of treating implicitly labeled instances different from explicitly labeled instances, even beyond the specific use case that we present here. Today, most extant research assumes a fixed label for implicitly labeled data (e.g., [27]), which potentially impedes performance and the development of explainable and trustworthy ML-based information systems. The results of our model training indicate that even with fewer, but high-quality samples, we are able to achieve similar prediction results. Based on our results, researchers should consider correcting mislabeled instances before publishing research data sets. If their goal, for example, is to publish a data set comparable to realistic real-world conditions, it may be inappropriate to correct mislabeled instances. If their goal is, however, to foster the development of trustworthy ML-based information systems for real-world environments, it may be important to correct mislabeled instances before publishing the data set.

For practice, our results highlight that practitioners should critically reflect on the label quality of public and private data sets that they want to use for supervised ML, especially in situations where trust in the ML models is essential. Lastly, our findings suggest that radiologists occasionally miss a disease when analyzing scans. Our findings, therefore, also support the deployment of AI-based medical image analysis systems to augment a radiologist's decision in real-time, which may improve the health outcome for patients.

6.3. Limitations and future research

We only evaluated our approach on the Cardiomegaly label and excluded labels representing

other diseases. With our selective focus on Cardiomegaly, we may have excluded correlations with other diseases present in the original study. Additionally, we were only able to evaluate a subset of around 30% of the patients present in the data set, since this was the size of our testing data set. Future research should therefore also apply our proposed system pipeline to other diseases in the CheXpert data set, as well as to other data sets for supervised ML. For this, potential data sets should have similar characteristics, such as a temporal dimension and uncertainty labels. Medical data sets often inherently exhibit these characteristics, however, our approach could also be applied to non-medical data sets. In general, incorporating distinctive data set characteristics into the analysis could provide a fruitful ground for further improvements in improving label quality. Another limitation pertains to the fact that we excluded implicit scans with the heuristic selection process, which may have resulted in the loss of some true positives. However, in doing so, we at the same time increased the precision of our proposed pipeline and, therefore, reduced the effort on radiologists, who may need to reevaluate flagged scans. Nevertheless, future research should further investigate the symbiosis of explicitly and implicitly labeled data regarding the benefits and risks of treating implicitly labeled data instances differently. Toward this end, other methods for handling implicit data should be explored, such as, for example, weighting implicit data instances less than explicit data instances. Lastly, we compressed the original X-ray scans for training purposes and may have lost relevant image information in the process. We chose this specific resolution to keep our training process within a reasonable time frame and stay as close as possible to the original CheXpert study.

7. Conclusion

In this research, we present a two-staged pipeline for the identification of potentially mislabeled instances in medical imaging data sets. Our results reveal that there are 7.4% of potentially mislabeled instances for the Cardiomegaly disease in the CheXpert data set, a data set that has received large attention from the computer vision research community. In developing the pipeline that we propose in this paper, we contribute to the current state of research, especially on label quality in supervised ML, in two ways. First, we show that the combination of ML-based approaches for the identification of labeling errors with data set-specific heuristic approaches serves to improve the precision with which labeling errors may be detected. Thus, it can reduce expert time required to identify mislabeled instances confidently. Second, we directly contribute to the CheXpert data set by identifying potentially mislabeled instances in the data set, thereby improving the data quality for a data set that is widely used in research. Our research is particularly interesting for researchers and practitioners who aim to contribute to the emerging field of trustworthy artificial intelligence, as the availability of high-quality data sets is a fundamental requirement [14].

8. Acknowledgements

We would like to thank Klaus Koch, who is a retired radiologist, for volunteering to participate in a blind study in order to verify that our system pipeline was capable of actually identifying mislabeled instances in the CheXpert data set.

References

[1] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya, *et al.*, "Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning," *arXiv:1711.05225v3*, 2017.

[2] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, *et al.*, "Mastering the game of go without human knowledge," *Nature*, vol. 550, no. 7676, pp. 354–359, 2017.

[3] K. D. Pandl, S. Thiebes, M. Schmidt-Kraepelin, and A. Sunyaev, "On the convergence of artificial intelligence and distributed ledger technology: A scoping review and future research agenda," *IEEE Access*, vol. 8, pp. 57075–57095, 2020.

[4] V. S. Sheng, F. Provost, and P. G. Ipeirotis, "Get another label? improving data quality and data mining using multiple, noisy labelers," in *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, p. 614–622, ACM, 2008.

[5] N. M. Müller and K. Markert, "Identifying mislabeled instances in classification datasets," in *2019 International Joint Conference on Neural Networks*, pp. 1–8, 2019.

[6] C. E. Brodley and M. A. Friedl, "Identifying mislabeled training data," *Journal of Artificial Intelligence Research*, vol. 11, pp. 131–167, 1999.

[7] J.-w. Sun, F.-y. Zhao, C.-j. Wang, and S.-f. Chen, "Identifying and correcting mislabeled training instances," in *Future Generation Communication and Networking*, vol. 1, pp. 244–250, IEEE, 2007.

[8] K. D. Pandl, H. Teigeler, S. Lins, S. Thiebes, and A. Sunyaev, "Drivers and inhibitors for organizations' intention to adopt artificial intelligence as a service," in *Proceedings of the 54th Hawaii International Conference on System Sciences*, 2021.

[9] J. Wiens, S. Saria, M. Sendak, M. Ghassemi, V. X. Liu, F. Doshi-Velez, K. Jung, K. Heller, D. Kale, M. Saeed, *et al.*, "Do no harm: a roadmap for responsible machine

learning for health care," *Nature Medicine*, vol. 25, no. 9, pp. 1337–1340, 2019.

[10] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak, B. van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, Dec. 2017.

[11] M. A. Gianfrancesco, S. Tamang, J. Yazdany, and G. Schmajuk, "Potential biases in machine learning algorithms using electronic health record data," *JAMA internal medicine*, vol. 178, no. 11, pp. 1544–1547, 2018.

[12] A. Rajkomar, M. Hardt, M. D. Howell, G. Corrado, and M. H. Chin, "Ensuring fairness in machine learning to advance health equity," *Annals of Internal Medicine*, vol. 169, no. 12, pp. 866–872, 2018.

[13] V. Eubanks, *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press, 2018.

[14] S. Thiebes, S. Lins, and A. Sunyaev, "Trustworthy artificial intelligence," *Electronic Markets*, 2020.

[15] P. Welinder and P. Perona, "Online crowdsourcing: rating annotators and obtaining cost-effective labels," in *2010 IEEE CVPRW*, pp. 25–32, 2010.

[16] Y. Yan, R. Rosales, G. Fung, and J. G. Dy, "Active learning from crowds," in *ICML*, 2011.

[17] C. Vondrick, D. Patterson, and D. Ramanan, "Efficiently scaling up crowdsourced video annotation," *International Journal of Computer Vision*, vol. 101, no. 1, pp. 184–204, 2013.

[18] L. Maier-Hein, D. Kondermann, T. Roß, S. Mersmann, E. Heim, S. Bodenstedt, H. G. Kenngott, A. Sanchez, M. Wagner, A. Preukschas, *et al.*, "Crowdtruth validation: a new paradigm for validating algorithms that rely on image correspondences," *International Journal of Computer Assisted Radiology and Surgery*, vol. 10, no. 8, pp. 1201–1212, 2015.

[19] A. G. S. de Herrera, A. Foncubierta-Rodrıguez, D. Markonis, R. Schaer, and H. Müller, "Crowdsourcing for medical image classification," in *Annual congress SGMI*, vol. 2014, 2014.

[20] L. Maier-Hein, S. Mersmann, D. Kondermann, C. Stock, H. G. Kenngott, A. Sanchez, M. Wagner, A. Preukschas, A.-L. Wekerle, S. Helfert, *et al.*, "Crowdsourcing for reference correspondence generation in endoscopic images," in *International Conference on MICCAI*, pp. 349–356, Springer, 2014.

[21] F. A. Schmidt, "Crowdsourced production of ai training data: How human workers teach self-driving cars how to see," tech. rep., Working Paper Forschungsförderung, 2019.

[22] A. Sorokin and D. Forsyth, "Utility data annotation with amazon mechanical turk," in *2008 IEEE CVPRW*, pp. 1–8, 2008.

[23] M. Murgia, *AI's new workforce: the data-labelling industry spreads globally*, 2019. Last accessed on 2020/07/15. *Available at* `https://www.ft.com/content/56dde36c-aa40-11e9-984c-fac8325aaa04`.

[24] G. Van Horn, S. Branson, R. Farrell, S. Haber, J. Barry, P. Ipeirotis, P. Perona, and S. Belongie, "Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection," in *IEEE CVPR*, pp. 595–604, 2015.

[25] S. Ørting, A. Doyle, A. van Hilten, M. Hirth, O. Inel, C. R. Madan, P. Mavridis, H. Spiers, and V. Cheplygina, "A survey of crowdsourcing in medical image analysis," *arXiv:1902.09159*, 2019.

[26] World Health Organization, "Communicating radiation risks in paediatric imaging," 2016.

[27] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya, *et al.*, "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 590–597, 2019.

[28] Y. Xie, M. Chen, D. Kao, G. Gao, and X. Chen, "Chexplain: Enabling physicians to explore and understanddata-driven, ai-enabled medical imaging analysis," *arXiv:2001.05149v2*, 2020.

[29] H. H. Pham, T. T. Le, D. T. Ngo, D. Q. Tran, and H. Q. Nguyen, "Interpreting chest x-rays via cnns that exploit hierarchical disease dependencies and uncertainty labels," *arXiv:2005.12734v1*, 2020.

[30] K. K. Bressem, L. Adams, C. Erxleben, B. Hamm, S. Niehues, and J. Vahldiek, "Comparing different deep learning architectures for classification of chest radiographs," *arXiv:2002.08991v1*, 2020.

[31] S. Minaee, R. Kafieh, M. Sonka, S. Yazdani, and G. J. Soufi, "Deep-covid: Predicting covid-19 from chest x-ray images using deep transfer learning," *arXiv:2004.09363v1*, 2020.

[32] J. P. Cohen, L. Dao, P. Morrison, K. Roth, Y. Bengio, B. Shen, A. Abbasi, M. Hoshmand-Kochi, M. Ghassemi, H. Li, *et al.*, "Predicting covid-19 pneumonia severity on chest x-ray with deep learning," *arXiv:2005.11856v3*, 2020.

[33] H. Amin and W. J. Siddiqui, "Cardiomegaly," in *StatPearls [internet]*, StatPearls Publishing, 2019.

[34] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," 2016.

[35] L. Prechelt, "Early stopping-but when?," in *Neural Networks: Tricks of the trade*, pp. 55–69, Springer, 1998.

[36] D. Hand and P. Christen, "A note on using the F-measure for evaluating record linkage algorithms," *Statistics and Computing*, vol. 28, pp. 539–547, May 2018.

[37] B. Frénay and M. Verleysen, "Classification in the presence of label noise: a survey," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, no. 5, pp. 845–869, 2013.

[38] T. C. Redman, "The impact of poor data quality on the typical enterprise," *Communications of the ACM*, vol. 41, no. 2, pp. 79–82, 1998.

[39] J. I. Maletic and A. Marcus, "Data cleansing: Beyond integrity analysis.," in *Iq*, pp. 200–209, Citeseer, 2000.

[40] L. Joseph, T. W. Gyorkos, and L. Coupal, "Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard," *American Journal of Epidemiology*, vol. 141, no. 3, pp. 263–272, 1995.

[41] A. Hadgu, "The discrepancy in discrepant analysis," *The Lancet*, vol. 348, no. 9027, pp. 592–593, 1996.

[42] I. Bross, "Misclassification in 2 x 2 tables," *Biometrics*, vol. 10, no. 4, pp. 478–486, 1954.