AI and Ethics (2021) 1:425–434
https://doi.org/10.1007/s43681-021-00049-0

ORIGINAL RESEARCH

Sensitivity of neural networks to corruption of image classification

Shimon Kaplan[1] · **Doron Handelman**[2] · **Amir Handelman**[1]

Received: 6 November 2020 / Accepted: 13 March 2021 / Published online: 23 March 2021

Abstract
Artificial intelligence (AI) systems are extensively used today in many fields. In the field of medicine, AI-systems are especially used for the segmentation and classification of medical images. As reliance on such AI-systems increases, it is important to verify that these systems are dependable and not sensitive to bias or other types of errors that may severely affect users and patients. This work investigates the sensitivity of the performance of AI-systems to labeling errors. Such investigation is performed by simulating intentional mislabeling of training images according to different values of a new parameter called "mislabeling balance" and a "corruption" parameter, and then measuring the accuracy of the AI-systems for every value of these parameters. The issues investigated in this work include the amount (percentage) of errors from which a substantial adverse effect on the performance of the AI-systems can be observed, and how unreliable labeling can be done in the training stage. The goals of this work are to raise ethical concerns regarding the various types of errors that can possibly find their way into AI-systems, to demonstrate the effect of training errors, and to encourage development of techniques that can cope with the problem of errors, especially for AI-systems that perform sensitive medical-related tasks.

Keywords Convolutional neural network · Artificial intelligence · Melanoma · Classification · Ethics

1 Introduction

Artificial intelligence (AI), which encompasses machine learning and deep learning, is viewed today as a leading field of development that is considered to affect many aspects of daily life including, for example in medicine [1], finance [2], and agriculture [3]. As the penetration of AI-systems becomes widespread, many systems are bound to rely upon it and to depend on its decision-making outcomes [4].

Today, AI-systems are already used in medicine, for example, for the recognition of anomalies in computed tomography (CT) images [5], and also for screening patients according to their illness severity [6]. The usefulness of such systems can be appreciated, for example, from * MERGEFORMAT [7] in which it was reported that a deep learning convolutional neural network (CNN) was able to recognize dermoscopic melanoma better than 58 dermatologists, 30 of which were specialized doctors. Along with this, it is important to make sure that such systems are reliable and have a minimal vulnerability to adversarial attacks and to accidental or other types of errors. Whereas in non-critical systems erroneous results generated by AI-systems may have little or practically no effect at all, in critical systems it is crucial for the results generated by AI-systems to be correct and reliable.

The main aim of this work is to increase awareness to the problem of the possibility of deliberately or inadvertently causing errors in training data. We also aim to encourage the development of tools and methods to prevent or at least minimize the effect of such a problem. We presume that deliberate corruption may result from hidden agenda, inherent racism or bias, political tendency or from various other reasons. By their nature, medical systems and applications are considered important and even critical and therefore avoiding or minimizing vulnerability to adversarial attacks is of particular interest [8].

Vulnerability to adversarial attacks may, for example, be evidenced from a study [9] by researchers from Harvard Medical School in which the researchers showed that adversarial examples in which pixels in images were modified in a way

Shimon Kaplan and Doron Handelman contributed equally

✉ Amir Handelman
handelmana@hit.ac.il

[1] Department of Electrical Engineering, Faculty of Engineering, Holon Institute of Technology, Holon, Israel

[2] Givatayim, Israel

that would seem like minimal noise to humans could manipulate deep learning systems that classify diabetic retinopathy from retinal images, pneumothorax from chest X-ray images, and melanoma from skin photos, and cause the systems to classify the pictures incorrectly.

In addition to attacks, errors could also stem from noise [10, 11], improper annotation [12], bias [13–15] and more. The omission of certain data types and lack of insertion of enough variety of data to enable the production of reliable results are also considered as causing errors [16, 17].

In the field of medicine, errors could affect all patients in a society or specific groups of patients, for example, according to their age, gender or race [18]. Coping with these kinds of errors is particularly important in these days of the global COVID-19 pandemic, where there is an outbreak of worldwide racism associated with the pandemic [19]. All of these types of errors, which may result from an AI-system that makes or assists in making medical diagnostic decisions such as the systems investigated in [5–7], may have adverse critical effects on health-care patients.

In this work, we show how an AI-system is influenced by labeling errors in the training data of the AI-system. Previous works have attempted to address an aspect of this issue regarding training neural networks based on unreliable labels [10, 11]. The issues investigated and discussed in the present work include the following: (1) the amount (percentage) of labeling errors from which a substantial adverse effect on the performance of the AI-systems can be observed, and (2) how unreliable labeling can be done in the training stage.

It is to be noted that errors in the training stage should not be the only matter of concern. Errors in the inference stage are of no less importance, and may possibly be of even greater significance. It is further to be noted that not only ethical issues should concern the public in connection with the analysis of medical images [20, 21], but also the interpretation of the results and reproducibility [22].

For the investigation of the effect of training stage errors on the performance of AI- systems we employed a convolutional neural network (CNN) on various types of datasets. First, in order to demonstrate how the investigated mislabeling works, we used a common dataset of cats and dogs. Then, we used a second CNN to perform a medical diagnostic task of distinguishing between malignant and benign skin moles. We mislabeled the images according to the gender information of the images, in order to demonstrate how bias based on gender can deliberately be generated. The second CNN serves as an example of a system that performs a critical task that, if affected by errors, may have adverse critical effects on health-care patients.

2 Description of error types and their relation to ethical issues

In the context of this paper, we refer to the term "error" in a broad sense to include any incorrect, inaccurate or unreliable or biased result, prediction, presentation, or classification resulting from the operation of an AI-system.

Although this work concentrates on training errors, as mentioned above, we believe it is useful to provide an overview of the various types (or sources) of errors that can possibly find their way into AI-systems. We categorize the various types of possible errors in AI-systems into two categories: training errors and inference errors. In respect of the category of training errors, some of the types of training errors include the following: (1) innocent/accidental errors, (2) intentionally-made errors, (3) omission-type errors, and (4) data errors.

We refer to innocent/accidental errors as errors that are made unintentionally, i.e., by mistake such as by a human labeler labeling some images incorrectly or tagging some elements in training dataset images incorrectly. This type of errors can inevitably occur and we consider it as unrelated to ethics.

We refer to intentionally-made errors as errors inserted on purpose by an adversary, such as by corruption and contamination of data or by deliberate mislabeling or insertion of incorrect, inaccurate, unreliable or biased data in an attack that is aimed to cause the AI-system to operate incorrectly or to otherwise influence the outcome of the AI-system for whatever reason. Such intentionally-made errors naturally create ethical problems. We presume that the adversary may insert such errors as a result of hidden agenda, inherent racism or bias, political tendency, conspiracy beliefs and anarchism, or from various other reasons. By inserting such intentionally-made errors, the adversary may change the outcome of an AI-system to produce results that meet his/her intentions.

There are many possible scenarios in which such cases may occur, depending on the application for which the AI-system is used. In one hypothetical example, an adversary may intentionally insert errors to cause incorrect training of an AI-system that is used for detecting susceptibility to a disease in a population to produce results showing that persons of one race are more susceptible to a certain disease than persons of another race. In such scenario, assuming the adversary has access to training data of the AI-system, the adversary may intentionally mislabel images of the training data such as to associate training images of many persons of the one race with genetic information and medical imaging results that are known to show susceptibility to the disease even though such association is incorrect. Following such mislabeling and training

thereafter, the AI-system is expected to produce, in the inference stage, racially-biased results, and people relying on such results may unjustly call for social distancing and separation between persons of the two different races.

In another hypothetical example, an adversary may intentionally insert errors to cause incorrect training of an AI-system that is used for mortgage approval to produce results that minimize the number of persons of one race whose mortgage requests are approved with respect to persons of other races and thus to reduce the number of habitants of the one race in a neighborhood. In such scenario, assuming the adversary has access to training data of the AI system, the adversary may choose training instances, for example according to typical family names of many of the persons of the one race, and then intentionally change attribute values that are used for classifying the instances to levels that cause denial of mortgage requests. Then, the adversary may train the AI-system again to prepare it for inference execution that is expected to produce results that deny mortgage requests of many persons of the one race. As AI-systems become more and more ubiquitous and their outcomes relied upon, the dangers of intentionally-made errors are expected to increase.

Regarding omission-type errors, we refer to them as errors caused by neglecting or omitting specific data types, by not referring to an entire span of possibilities, or by using imbalanced training data (class imbalance). For example, an AI-system that is intended to perform face recognition is bound to make determination mistakes if it does not include, in its training dataset, images with attributes of a variety of persons of multiple human races and of persons of different genders and different ages of the multiple human races and if the training dataset is severely imbalanced with respect to many of the attributes. This type of errors is directly linked with ethical issues although we presume that such errors are not intentional (if they are intentional, then such errors are considered and referred to as part of the intentionally-made errors).

Omission-type errors can occur for various reasons including, for example, lack of sufficient diversity of training data, negligence, lack of awareness, and disregarding a failure of an AI-system [17]. We believe that education may be useful to prevent or minimize such errors and to raise awareness to the possibility of the existence of such errors, and we provide hereinafter some suggestions for avoiding or at least minimizing the possibility of existence of such omission-type errors. A famous example to omission-type errors are biased facial-recognition systems [23].

Regarding data errors, we refer to them as errors that may be found in the training dataset, which result from natural phenomena, such as noise [11] or from instabilities of deep learning systems with respect to image reconstruction/classifications due to tiny perturbation, such as patients small movements [24]. This type of errors can inevitably occur and we consider this type as unrelated to ethics, but nevertheless, we recommend cautious selection of sources for the collection of training data to avoid or at least reduce the number of data errors therein.

In respect of the category of inference errors, some of the types of inference errors include approach-induced errors and technical errors.

We refer to approach-induced errors as errors that are system-wide and may be caused by an incorrect or inadequate approach that is adopted for solving a problem using an AI-associated system. Such approach-induced errors may result from:

1. Incorrect or incomplete definition of the problem and/or the solution sought
2. Poor adaptation of the problem to a solution based on AI-systems
3. Use of an AI-system that is inadequate for solving the specific problem
4. A bias inherent in the *definition* of the problem.

With respect to bias, we note that bias can appear in the training phase or in the inference phase, or in both. Bias is typically associated with applications that involve humans and in the training phase, it can be expressed, for example, by mislabeling one or more training data images whereas in the inference phase it can be caused as part of an underlying algorithm. Bias can be based, for example, on gender, race, age, income, or a combination thereof.

In the inference stage bias can be direct or indirect. A direct bias may, for example, result from a design decision to refer differently to different genders. An indirect bias may, for example, result from knowingly or unknowingly taking into account a parameter, such as zip code, from which differentiation based on income can be made with high certainty.

As for technical errors, we refer to them as errors resulting from technical problems, or from intentional or unintentional mistakes, inaccuracies or omissions in design, algorithms, programming, and implementation, or from any combination thereof.

At least some of the above-mentioned errors may take one or several forms. For example, a data error may be an error caused by noise or an error caused by speckle, an intentionally-made error may be an error caused by an adversarial patch or an error made by incorrect labeling, and so on.

It is to be noted that each of the above-mentioned list of training errors and list of inference errors is not necessarily exhaustive, and other types of errors may also occur.

In this paper, we concentrate on the category of training errors and particularly on intentionally-made errors inserted

by an adversary. Furthermore, in the category of training errors we focus on labeling errors.

3 Related works

Noisy training data and particularly labeling errors have been the subject of many studies. Brodley and Friedl [25] use a set of learning algorithms to create classifiers that serve as noise filters. They evaluate single algorithm, majority vote, and consensus filters and show that filtering significantly improves classification accuracy for noise levels up to 30%.

Bekker and Goldberger [11] propose an algorithm for training neural networks based solely on noisy data. They introduce an extra noise layer by assuming that observed labels are created from true labels by passing through a noisy channel, and show that they can learn the noise distribution from noisy data without using any clean data. Goldberger and Ben-Reuven [26] model noise by an additional softmax layer that connects the correct labels to the noisy ones. Dgani, Greenspan and Goldberger [10] use a noisy channel in their neural-network training strategy which is based on unreliable human annotation.

Guan et al. [27] presents an approach for identifying and eliminating mislabeled training instances by using unlabeled instances to aid detection of the mislabeled training instances, and Khoshgoftaar et al. * MERGEFORMAT [28] addresses the combined effects of class imbalance and labeling errors.

The works mentioned above deal with noisy data, which is usually related to the data collection process [10], * MERGEFORMAT [26]. As such, the works can probably be suitable for the type of training errors that we refer to as data errors, but not necessarily for the intentionally-made errors of the training errors category that are not expected to be modeled as noise.

Additionally, in respect of class imbalance when it is intentionally-made, that is, when it is part of the intentionally-made errors, an adversary may decide to force class imbalance or to change important class examples that form the basis for oversampling and undersampling so as to increase the amount of errors in data instances and therefore such class imbalance cannot be treated by data sampling techniques as the class imbalance in collected data.

Furthermore, it can also be assumed that in respect of intentionally-made errors, the adversary may insert them after filters and other algorithms that are intended to improve classification accuracy and class balance have already been used. Accordingly, the works mentioned above do not solve ethical and functional problems associated with some types of training errors, and further actions, procedures and tools are needed to assist in coping with such ethical and functional problems.

4 Methods

The work was implemented in Python™ with Keras and TensorFlow in a Windows 10 environment and was run on the Nvidia® GeForce® GTX 1060 graphics processing unit (GPU).

4.1 CNN construction – dogs-and-cats dataset

The dataset was taken from Kaggle (https://www.kaggle.com/chetankv/dogs-cats-images) and included 25,000 images of two classes: dogs and cats. The dataset was divided into a training set for network training, which included 80% of the data (20,000 images in total, 10,000 for each class), and a test set for testing network results, which included 20% of the data (5000 images in total, 2500 for each class). Our expectation was to reach an accuracy higher than 80% for our initial baseline before we start to gradually mislabel the images and to re-run the model (Fig. 1).

Our network was produced using a classic two convolutional block topology. Each layer had fixed parameters and hyper parameters that were set as Keras's default values. The network architecture was as follows (Fig. 2):

The parameters for the network are:

Input Layer

RGB (red, green, blue) images of 202×180 size allocated to the training set and the test set.

Hidden Layers

Fig. 1 Examples of images from the Kaggle dataset of: (**a**) cats, and (**b**) dogs. The numbers at the top left sides of the images are the probabilities that the images are "cat"

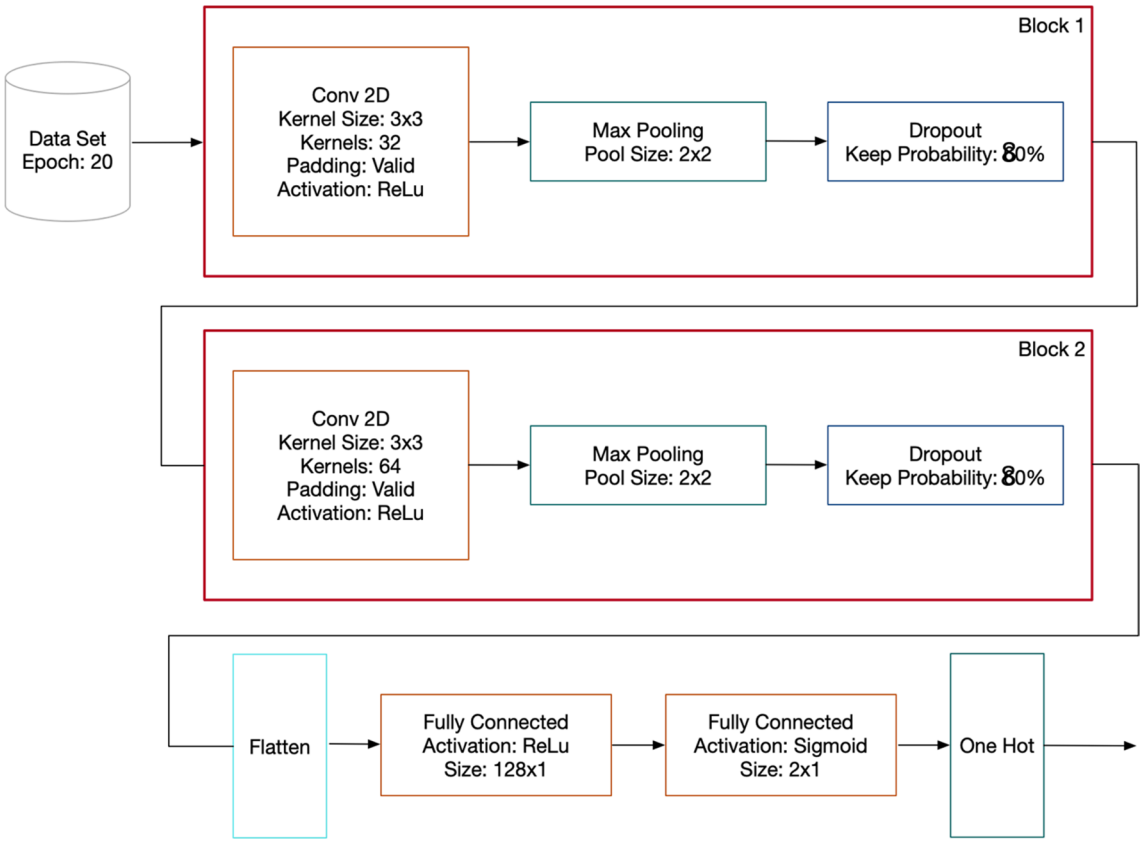

Fig. 2 Network architecture

- First layer
  - o Convolution layer – 32 filters of 3 × 3 size, 'ReLU' activation
  - o Maxpooling – of 2 × 2 size
  - o Dropout – 60% (keep probability)
- Second layer
  - o Convolution layer – 64 filters of 3 × 3 size, 'ReLU' activation
  - o Maxpooling – of 2 × 2 size
  - o Dropout – 80% (Keep probability)

Classification layers:

- Flattening layer
- Compression layer (Dense) 128 × 1 size, 'ReLU' activation
- Compression layer (Dense) 2 × 1 size, 'Sigmoid' activation
- One hot

Random seed: initialized to 2019.

Number of Epochs (an epoch is one cycle through the complete training dataset): 20.

Loss function: Binary_Crossentropy.

Optimization method: Adam, learning rate default: 0.01.

We loaded images that passed real-time data augmentation by the ImageDataGenerator class in the Keras library and were looped over in batches. The augmentations performed on the training set were as follows:

- Rescale – normalization by 1/255
- Shear_Range – diagonal shearing of the image up to 20%
- Zoom_Range – increasing/decreasing image size up to 20%
- Horizontal_Flip – flipping on the horizontal axis
- Width_Shift_Range – left/right image shifting up to 10%
- Height_Shift_Range – up/down image shifting up to 10%
- Rotation_Range – image rotation up to 45 degrees

4.2 CNN construction – melanoma PH2 dataset

In this dataset (https://www.fc.up.pt/addi/ph2%20database.html) there are a total of 200 images, which consist of 40 images classified as malignant Melanoma and 160 images classified as 'Common Nevus' and 'Atypical

Nevus' that can be summed up as a 'Non-Melanoma' class. Some samples of images classified as 'Melanoma' and 'Non-Melanoma' are shown in Fig. 3.

Using basic data-augmentation including horizontal and vertical flip and image noise, we increased the total number of images from 200 to 12,800, where 2560 were classified as 'Melanoma' and 10,240 were classified as 'non-Melanoma'. We used these images for testing with the *same* network and parameters listed in the dogs-and-cats section above and the same balance-corruption method.

4.3 CNN Construction – melanoma ISIC-ARCHIVE dataset

We reached the dataset of ISIC (International Skin Image Collaboration) that includes images of skin cancer, with particular emphasis on melanoma. This dataset includes pictures of malignant and benign skin moles with different resolutions and attributes (https://www.isic-archive.com/#!/topWithHeader/wideContentTop/main).

As opposed to the PH2 dataset, ISIC-ARCHIVE offers a couple of thousands of images with larger meta-data information, including gender and age. Using this dataset we tried to simulate a possible intentional-error insertion situation where images are mislabeled as a result of discrimination, for instance – a gender-based discrimination. We chose images with the meta-data's gender 'female' and gradually mislabeled the images to see how the accuracy of the model is affected.

In this dataset, the original number of male-related images was 6255 and the original number of female-related images was 4727. Using basic data-augmentation including horizontal and vertical flip, and rotation at 90°, 180° and 270°, we increased the total number of female-related images to 23,635 and the total number of male-related images to 31,275. Then, 90% of the total number of images were taken for training, and 10% of the total number of images were taken for testing.

4.4 The balance-corruption method

After reaching a reasonable functioning model with accuracy higher than 80%, we started to mislabel the images of our training dataset using two parameters: mislabeling balance, and corruption.

The mislabeling balance is a parameter that expresses from which type of dataset the corruption is to be made. For example, when the mislabeling balance value is −1, it means that 100% of the images in category A (such as cats or benign non-melanoma) are to be used as an image bank from which a particular number of labels (defined by the corruption parameter) are to be corrupted, and none (0%) of the images in category B (such as dogs or malignant melanoma) are to be corrupted. When the mislabeling balance value is −0.5, it means that 75% of the images in category A and 25% of the images in category B are to be used as the image bank from which a particular number of labels (defined by the corruption parameter) are to be corrupted. When the mislabeling balance value is 0, it means that 50% of each category are to be used as the image bank from which a particular number of labels (defined by the corruption parameter) are to be corrupted. The definitions of mislabeling balance and corruption are summarized in Table 1.

For each mislabeling balance value group, the corruption parameter determines the total amount (in percentage values) of images in the group to be randomly selected and mislabeled. Note that the corruption value ultimately yields the final percentage of corrupted images. As an example, which refers to the cats-and-dogs dataset, when the mislabeling balance value is 0 and the corruption value is 10%, it means that 10% of the images from the 50% image bank of each category are to be corrupted. Therefore, if we have 5000 cat images and 5000 dog images and we set aside randomly 2500 cat images and 2500 dog images as image banks for corruption, 10% corruption means that we corrupt 250 cat images and 250 dog images. When the mislabeling balance value is − 1 and the corruption value is 10%, it means that 10% of the images from the entire (100%) image bank of cats are to be corrupted and no dog images are to be corrupted.

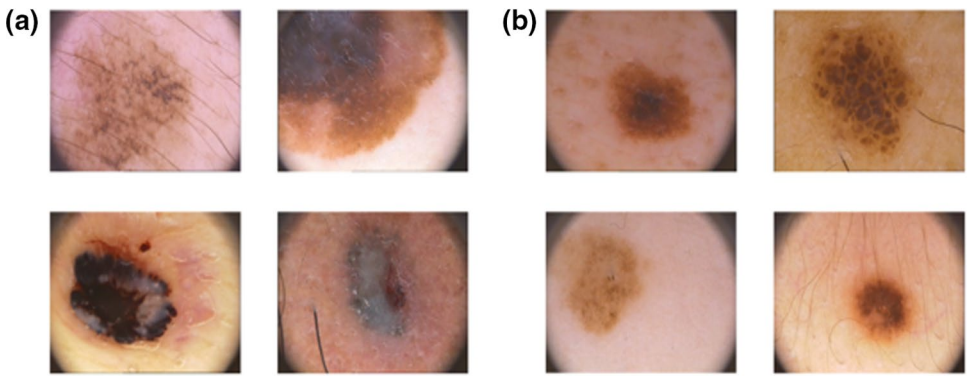

Fig. 3 Examples of images from the PH2 dataset: (**a**) Melanoma, (**b**) Non-Melanoma

Table 1 Definition of mislabeling balance parameter for two category dataset

Mislabeling balance value	Image category (percentage)	Num. of images in category A	Num. of images in category B
−1	100% category A 0% category B	10,000 (cats) 10,240 (non-melanoma)	0 (dogs) 0 (melanoma)
−1/2	75% category A 25% category B	7500 (cats) 7680 (non-melanoma)	2500 (dogs) 640 (melanoma)
0	50% category A 50% category B	5000 (cats) 5120 (non-melanoma)	5000 (dogs) 1280 (melanoma)
+1/2	25% category A 75% category B	2500 (cats) 2560 (non-melanoma)	7500 (dogs) 1920 (melanoma)
+1	0% category A 100% category B	0 (cats) 0 (non-melanoma)	10,000 (dogs) 2560 (melanoma)

*The table is related to dogs-and-cats and PH2 datasets

Therefore, if we have 10,000 cat images, the entire 10,000 cat images are set aside for corruption, and 10% corruption means that we corrupt 1000 cat images.

We used the balance-corruption method in a somewhat different way for the ISIC-ARCHIVE dataset. We wanted to test a scenario where only one category (one type of gender) is being mislabeled as a result of malicious pre-intended mislabeling. For example, if some benign image is related to a female, corruption of this image means a change of the classification to malignant (and vice versa). To achieve this, we took a portion of the images that are related to this gender category (we used only the female-related category that included 21,272 images after augmentation), and then we gradually corrupted these female-related images. Corruption in this case is still the percentage out of that amount which is being manipulated.

As an example, we started by taking only 10% out of the 21,272 female-related images (that is, 2127 images) and then we corrupted 10%, 20%, 30% and 40% out of these 2127 images (that is, 213, 425, 638, and 851 images, respectively). Next, we took 20% out of the 21,272 female-related images (that is, 4254 images) and then we corrupted 10%, 20%, 30% and 40% out of these 4254 images (that is, 425, 851, 1276, and 1702 images, respectively). This continued until we took 100% of the female-related images.

5 Results

As mentioned above, we aim to promote discussion regarding how to detect and cope with deliberate insertion of errors in training data, and particularly mislabeling of training images due to various reasons, such as hidden agenda, inherent racism or bias, or political interest.

We start by showing how such mislabeling can occur. The first example is mislabeling a simple dataset of dogs and cats. Using the method described above, we show (Fig. 4) the dependence of accuracy on corruption for different mislabeling balance values.

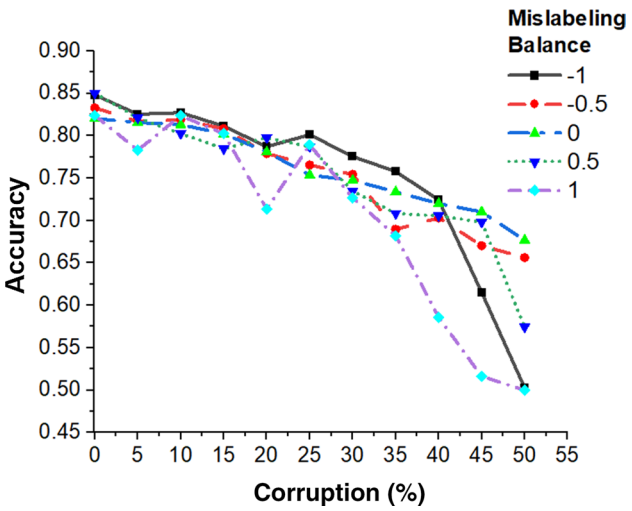

Fig. 4 Dogs-cats image accuracy vs. corruption for different mislabeling balance values

In Fig. 4 that is related to the dogs-and-cats dataset, we can see that in all tested combinations of mislabeling balance and corruption, the model maintains a reasonable performance up to the 25% corruption region. Further on from there, the model's accuracy drops drastically from the 82–85% baseline.

While an application that distinguishes between dogs and cats is most likely non-critical, and therefore damaging the classification of images in the training stage is not expected to have severe implications, such damaging may be disastrous in a critical application, such as a medical application.

In Fig. 5 that is related to the PH2 dataset, we can see that accuracy is more dependent on mislabeling balance, and as in the previous section, even a slight corruption of 5% can reduce accuracy by up to 9% (from 0.9 to 0.825) depending on mislabeling balance value. Here, a major reduction

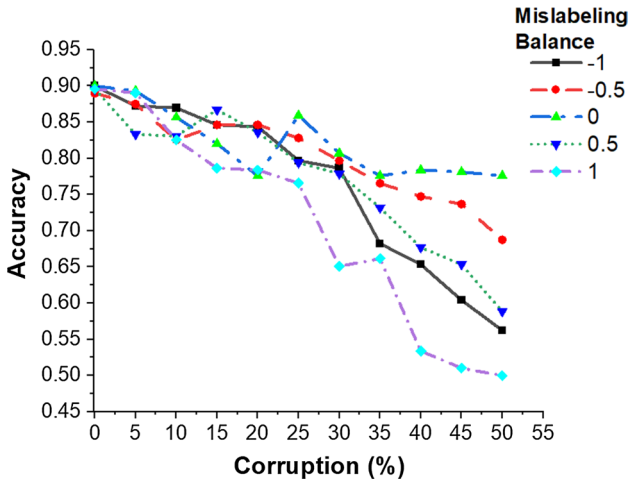

Fig. 5 Melanoma PH2 Dataset image accuracy vs. corruption for different mislabeling balance values

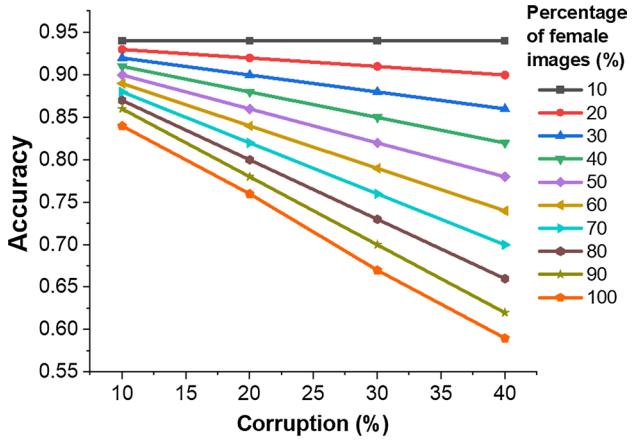

Fig. 6 Accuracy vs. image corruption according to a gender-bias in the Melanoma ISIC-ARCHIVE for different mislabeling balance values

occurs at lower percentage values of corruption, most likely due to the small number of images in the dataset.

We now focus on corruption according to a gender bias in relation to the ISIC-ARCHIVE dataset, that is, for images related to one gender we deliberately mislabeled some of the non-Melanoma images as Melanoma images.

As shown in Fig. 6 that is related to the ISIC-ARCHIVE dataset, when only a small portion of the female-related images is corrupted the corruption has little effect on the accuracy. This can be observed from the plateau graph when we take only 10% out of the female-related images and corrupt up to 40% of those images. This result is most likely due to the small amount of images (up to 851 images) that were corrupted. When the corruption is done on a larger percentage of female-related images, starting from around 20%, the corruption affects the accuracy dramatically. The graphs are close to linearity most probably since we took only 4 corruption values, and up to 40% only.

6 Discussion

Among the many types of errors that can find their way into an AI-system, intentional corruption of training images is one type of errors that should be of particular concern. Such intentional corruption may take many forms including, for example, mislabeling of images and insertion of false data or patches in images.

In the methods of this work, we concentrate on intentionally-made errors in the form of labeling errors caused by an adversary. We assume that we cannot predict any of the following: the amount of images the adversary decides to mislabel; a stage at which mislabeling occurs, namely whether the adversary decides to mislabel the images collected or the images after use of filtering or other algorithms that are intended to improve classification accuracy and class balance; and the distribution of mislabeled images across the different classes of the training dataset and throughout the entire training dataset, namely whether the adversary mislabels different numbers or different percentages of images in different classes, and which images of each class are actually mislabeled.

In order to demonstrate the effect of intentional mislabeling of training images, in view of such lack of predictability, we used a model that analyzes accuracies of the AI system in the presence of labeling errors, that is, the performance of the AI system when the images that are intentionally mislabeled form part of the training dataset and the AI system is trained with this training dataset. Accordingly, in this model each labeling error has the same weight and effect regardless of the reason for creating it, and therefore it makes no difference, for the purpose of this model, whether the adversary inserts a labeling error due to bias (e.g., an undesired gender attribute value), or due to inherent racism (e.g., an undesired skin color attribute value), or due to any other reason. This, therefore, makes the analysis regarding the exemplary CNNs mentioned above more generally applicable and more general in nature and suitable for dealing with a variety of intentionally-made errors associated with different attributes of images.

As for the amount and distribution of the labeling errors, it can be assumed that the adversary may prefer for his/her mislabeling to be undetected and therefore not to overdo the corruption and to spread the labeling errors over many parts of the training dataset. Therefore, it is important to develop and adopt tools and methods that detect even small amounts of intentionally-made errors and prevent or at least minimize the effect of such errors.

In all datasets investigated in this work severe reductions in accuracy for all mislabeling balance values occurred from 35% corruption. This means that to maliciously mislabel the images, an adversary needs to mislabel at least 35% of the images (without such mislabeling being detected) in order to get errors that can severely damage the classification system.

Looking at small amounts of corruption, however, we can see, for example, that a 5% corruption reduces accuracy by approximately 2%. Such a drop in accuracy may be considered insignificant for non-critical applications (such as the application that distinguishes between dogs and cats), but for a critical application (such as a medical application) such drop in accuracy can be significant, for example as mentioned in [29] (where it is argued that even 1% poor-quality data can impact the performance of the AI system), and especially when it is known that there are one billion radiologic examinations every year * MERGEFORMAT [30].

One factor that is noted from the results mentioned above as influencing the level of accuracy and accuracy sensitivity of the AI-system in the presence of training labeling errors is the size of the training dataset. The AI-system can better cope with errors when the training dataset is large. In the particular scenario of the dogs-and-cats dataset, a large number of images in the dataset successfully compensated for the forced mislabeling. However, this behavior cannot be guaranteed if a much smaller dataset is available with only a few hundreds of images, or even less. Therefore, when a developer claims to use AI in a system developed thereby, one of the first issues to look at and evaluate is the size of the training dataset the developer uses.

In addition, we showed how easily one can manipulate labeling in the training phase. Such manipulation is bound to affect the AI-system, and thus it is dangerous to treat an AI-system as a "black-box". Since image labeling may be easily outsourced, a developer needs to be aware of the possibility of labeling manipulation.

In light of our findings, we provide hereinafter some suggestions for preventing/avoiding or at least minimizing the possibility the intentionally-made labeling errors. First, we recommend that image labeling be performed by at least two unrelated entities and their labeling outcomes be compared to identify and correct discrepancies before their use in AI training. Second, we recommend that AI developers should use collaborations of individuals of many races, religions, genders, and even nations. We believe that companies that publish the anonymous profiles of the developers' background can give more confidence to people that use AI-systems. Third, methods to spot mislabeling of images should be continuously developed. For example, the number of epochs and revisions used in a CNN should be monitored to ensure that no unwanted reruns or revisions have been made after confirmation of a version. Further, for example, the output of the AI-system should be tested by several independent professionals in a field (this can be viewed as practice-based approaches). Additionally, it is important to verify the diversity of the data inputted into AI-systems (in medical applications – images from different genders, races, etc.), and any anomalies should be searched and examined for each specific data type (this can be viewed as data-based approaches).

7 Conclusions

In this work, we investigated the level of accuracy and sensitivity of an AI-system in the presence of labeling errors in the training dataset. The results obtained in this work are particularly discussed in connection with intentionally-made errors. Such investigation is considered important because using AI-systems with errors in production can negatively affect decision-making. The investigation was carried out by selectively inserting labeling errors in three types of datasets (cats-and-dogs, Melanoma PH2, Melanoma ISIC) according to different values of a corruption parameter and different values of a mislabeling balance parameter, and then measuring the accuracy of the AI-system for every value of these two parameters. It was demonstrated that even a slight corruption can reduce the accuracy depending on mislabeling balance parameter value. With the increased use of AI-systems in various fields, and especially in medical fields, the goal of this work is to raise questions regarding the accuracy of these AI-systems and to warn against treating AI-systems as black-boxes that always produce results that are correct and unquestionable.

Declarations

Conflict of interest The authors declare that there is no conflict of interest.

Code availablity The code is available from https://github.com/siomak/data_set_corruption

References

1. Mintz, Y., Brodie, R.: Introduction to artificial intelligence in medicine. Min. Inv. Ther. All. Tech. **28**, 2–73 (2019)
2. Bredt, S.: Artificial Intelligence (AI) in the financial sector—Potential and public strategies. Front. AI. **2**, 3 (2019)
3. Kirtan, J., Aalap, D., Poojan, P., Manan, S.: A comprehensive review on automation in agriculture using artificial intelligence. AI. Agr. **2**, 1–2 (2019)
4. Jarrah, M.H.: Artificial intelligence and the future of work: human-AI symbiosis in organizational decision making. Bus. Horiz. **61**(4), 577–586 (2018)

5. Prevedello, L.M., Erdal, B.S., Ryu, J.L., Little, K.J., Demirer, M., Qian, S., White, R.D.: Automated critical test findings identification and online notification system using artificial intelligence in imaging. Radiology **285**(3), 923–931 (2017)
6. Paiva, O.A., Prevedello, L.M.: The potential impact of artificial intelligence in radiology. Radiol. Bras. **50**(5), V–VI (2017)
7. Haenssle, H.A., et al.: Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to dermatologists. Ann. Oncol. **29**(8), 1836–1842 (2018)
8. Kaissis, G.A., Makowski, M.R., Rückert, D., et al.: Secure, privacy-preserving and federated machine learning in medical imaging. Nat. Mach. Intell. **2**, 305–311 (2020)
9. Finlayson, S.G., Bowers, J.D., Ito, J., Zitrain, J.L., Beam, A.L., Kohane, I.S.: Adversarial attacks on medical machine learning. Science **1**, 1287–1289 (2019)
10. Dgani, Y., Greenspan, H., Goldberger, J.: Training a neural network based on unreliable human annotation of medical images, IEEE 15th International Symposium on Biomedical Imaging, Washington, DC, 39–42 (2018)
11. Bekker, J., Goldberger, J.: Training deep neural-networks based on unreliable labels, Proceedings of the IEEE international conference on acoustics, speech and signal processing (ICASSP), 2682–2686 (2016)
12. Xia, F., Yetisgen-Yildiz, M.: Clinical corpus annotation: challenges and strategies. Proceedings of the Third Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM'2012) in conjunction with the International Conference on Language Resources and Evaluation (LREC), Istanbul, Turkey, (2012)
13. Gianfrancesco, M.A., Tamang, S., Yazdany, J., Schmajuk, G.: Potential biases in machine learning algorithms using electronic health record data. JAMA Intern. Med. **178**(11), 1544–1547 (2018)
14. Hutson, M.: It's too easy to hide bias in deep-learning systems. IEEE Spect. **1**, 2–19 (2021)
15. Challen, R., Denny, J., Pitt, M., et al.: Artificial intelligence, bias and clinical safety. BMJ Qual. Saf. **28**, 231–237 (2019)
16. Qayyum, A., Qadir, J., Bilal, M., Al-Fuqaha, A.: Secure and robust machine learning for healthcare: a survey. Preprint at https://arxiv.org/abs/2001.08103 (2020)
17. Geis, J.R., Brady, A.P., Wu, C.C., et al.: Ethics of artificial intelligence in radiology: summary of the joint European and North American Multisociety Statement. Radiology **293**(2), 436–440 (2019)
18. Strickland, E.: Healthcare algorithms show racial bias. IEEE Spect. **8**, 6–7 (2020)
19. Zannettou, S., Baumgartner, J., Finkelstein, J., Goldenberg, A.: Weaponized information outbreak: a case study on COVID-19. Bioweapon Myths and the Asian Conspiracy Meme (2019)
20. D'Antonoli, T.A.: The ethical considerations for artificial intelligence: an overview of the current radiology landscape. Diagn. Interv. Radiol. **26**, 504–511 (2020)
21. Larrazabal, A.J., Nieto, N., Peterson, V., Milone, D.H., Ferrante, E.: Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. PNAS **117**(23), 12592–12594 (2020)
22. Maier-Hein, L., Eisenmann, M., Reinke, A., et al.: Why rankings of biomedical image analysis competitions should be interpreted with care. Nat. Commun. **9**, 5217 (2018)
23. Castelvecchi, D.: Is facial recognition too biased to be let loose? Nature **587**, 347–349 (2020)
24. Antun, V., Renna, F., Poon, C., Adcock, B., Hansen, A.C.: On instabilities of deep learning in image reconstruction and the potential costs of AI. Proc. Nat. Aca. Sci. **117**(48), 30088–30095 (2020)
25. Brodley, C.E., Friedl, M.A.: Identifying mislabeled training data. J. Arti. Intell. Res. **131–137**, 11 (1999)
26. Goldberger, J., Ben-Reuven, E.: Training deep neural-networks using a noise adaptation layer. ICLR **6**, 6 (2017)
27. Guan, D., Yuan, W., Lee, Y.K., Lee, S.: Identifying mislabeled training data with the aid of unlabeled data. Appl. Intell. **35**, 345–358 (2011)
28. Khoshgoftaar, T.M., Hulse, J.V., Napolitano, A.: Supervised neural network modeling: an empirical investigation into learning from imbalanced data with labeling errors. IEEE Trans. Neu. Net. **21**(5), 813–830 (2010)
29. https://www.itnonline.com/content/ai-algorithm-detects-difficult-read-medical-images
30. Brady, A.P.: Error and discrepancy in radiology: inevitable or avoidable? Insights Imaging **8**, 171–182 (2017)