Assessing the Security of Personal Data in Large Scale chest X-Ray Image Screening

Vassili Kovalev
Department of Biomedical Image Analysis
United Institute of Informatics Problems
Belarus National Academy of Sciences
Minsk, Belarus
vassili.kovalev@gmail.com

Abstract— **This paper is related to the problem of security of personal data in the context of massive screening of the population. We attempt to assess the probability of identification of particular individuals by their chest images stored together with other private data in large-scale image databases. It was supposed that the attacker have X-Ray image of target subject but taken several years earlier at different scanning conditions and uses it as a key. A total of 90,000 images of 45,000 subjects were sampled from a database containing 1,909,000 records. The study groups were fully balanced by both age and gender. Image features were derived using 3 different CNNs including EfficientNet-B0, EfficientNet-B0-V2, and BiT-S R50x1. Results of searching correct image of a pair for all 45,000 people were presented in form of the fraction of correct answers in the Top-N most similar while N runs from 1 (correct answer on the first position) up to 80. It was found that (a) EfficientNet-B0 produces the best image features among the three CNNs being examined. (b) The fraction of correct identifications of subjects that is the right answers appeared on the first position was about 14% whereas the percentage of correct results in Top-80 achieved 33%. (c) The chances to be identified are significantly higher in female subjects compared to males and higher in young subjects compared to the aged ones.**

Keywords—Personal data security, Medical images, Searching similar images, Convolutional neural networks.

I. INTRODUCTION

It is known that computer-assisted screening of the population is an important way of timely discovering of lung diseases [1] as well as some other abnormalities in chest [2]. One of the results of such a screening is that the resultant image database contains large fraction of normal cases whereas different kinds of abnormalities represented proportional to their natural incidence rates in the population. Lately, under condition of a broad use of Deep Learning technologies for computerized disease diagnosis, there are clear signs of growing concerns related to the security of personal data [3–5]. Both generally positive trends including the permanently growing image-based health care processes as well as the implementation of massive population screening measures are inevitably paired with the growth of the doubts regarding the security of patients' personal data. These worries are ranged from the ethical and aesthetical concerns associated with the disclosure of personal anatomy and health condition details and going up to the risk of possible financial losses and even to the complete destruction of private life. In this work, we are focusing on the scenario of the use of a patient's medical image as a key for identification of target person in large databases of Picture Archiving and Communication Systems (PACS) containing images of the same medical modality. We do not suppose that among the images stored in the database there is one, which is identical to the image available for the malicious attacker. That would be a rather simple case. Instead, we only admit that images were acquired from the same person at different time points and under different image acquisition conditions. We are posing the problem of person identification as a sort of similarity retrieval. More specifically, for every pair of different images of the same person we are going to estimate quantitatively the chances to find out the second image using the first one as a key for searching in a large database. In order to achieve this, we employed large set of chest X-Ray images. It is not supposed that among the images stored in the database there is one, which is identical to the image available for the malicious attacker. Instead, we only admit that they were acquired from the same person. In this study, we attempt to assess the probability of identification of particular individuals in large-scale image databases resulted from the screening of lung diseases. With that in mind, we are not only addressing the issue of likelihood of identifying of database records by querying chest X-Ray images as such. We also take into account the age range and gender of each subject. This is because it is a well-known fact, accepted by the medical image analysis community, that both age and gender play a very important role in computerized image analysis. More related details can be found, for example, in our previous work on lung image analysis [6], which was performed on a dataset of very similar images consisting of 188,000 items sampled from the master image database containing several millions of records. Image features used for computing the degree of similarity were extracted with the help of Convolutional Neural Networks (CNNs) of different architectures. Thus, in general, the results of this paper may be viewed as a contribution to the experimental materials helping to improve the security of medical information systems as well as the response to the increasing attention to the data security problems. Recently, such a trend can be observed in different countries of the world which permanently continuing the process of imposing additional restrictions in the regulations related to the safety of personal data.

II. MATERIALS: THE INPUT IMAGE DATA

The image sources. All the chest X-Ray image data used in this study were the natively-digital X-Ray scans resulted from massive screening of the population for early diagnosis of lung diseases as well as for detection of cerebrovascular abnormalities and pathological changes of the skeleton (see [2] and free web-based diagnostic services [7] for more details).

Creating the local image repository. According to the main goal of this exploratory study, the input image dataset was combined of X-Ray images of subjects who passed the digital X-Ray examination two or more times. The number of corresponding database records available for experimentation has been amounted up to 1,909,000 items. Each record corresponds to a single digital chest X-Ray image. The records

include information on patients' age and gender as well as the textual radiological reports. Textual reports were parsed by a separate project which ends up with the categories of health condition and classification of the original chest image dataset to subgroups. In terms of the frequency of presence of different age groups, genders, types of body constitution and other factors, the source images represent the natural appearance of these subgroups in the population. The main steps of the procedure of creating fully balanced study groups used in this work are described below.

Step-1. Selection of healthy subjects scanned two or more times. For each of them, taking two X-Ray scans at random with no matter how many of them are available. The procedure resulted in approximately 140,000 images of 70,000 subjects scanned at two different time points. The information about corresponding time gaps between image acquisition events was not available because all the scans were fully anonymized including the scanning date and place.

Step-2. Given that subjects of different age are represented in reasonably different quantities, we have chosen to sample images of subjects between 18, what is required by actual national law, and 60 years of life.

Note that the number of persons aged 60 above years who wish to pass the screening examination drops down reasonably quickly. Clearly, the above decision was a tradeoff between the aspiration to have the study group as large as possible, on the one hand and the strong pre-requirement for creating well-balanced image dataset on the other.

Step-3. In addition, we also applied a very strong image sub-sampling requirement of absolutely equivalent representation of both female and male genders in each group and sub-group.

As a result of application of the above three criteria, we end up with about 100,000 images in total.

Step-4. In order to provide certain degree of separation between the age groups, we have decided to create 3 age groups of 10 years life span each with the two age gaps of 5 years in between.

Applying all the above requirements resulted in 3 perfectly balanced age sub-groups ("classes") of healthy subjects which are described in Table 1.

The naming conventions. It should be remembered that each of the above 45,000 subjects aged 18-57 was scanned twice. Thus, the total number of X-Ray images was 90,000. The resultant C1, C2, and C3 age classes conditionally called here as Young, Middle-aged, and Aged. These are conditional labels assigned purely for distinguishing our study sub-groups which do not have any direct biomedical interpretation. It must also be realized that sub-dividing people to certain age groups is rather complicated problem. It varies substantially and depends on the country, the minimum age allowed to starting legal working, the actual national retirement law, corresponding legal differences associated with gender, the existing national traditions, etc.

TABLE I. IMAGE STUDY GROUPS

Group acronym	Age range	Total people	Number of males	Number of females
C1, Young	18-27	15,00	7,500	7,500
C2, Middle	33-42	15,00	7,500	7,500
C3, Mature	48-57	15,00	7,500	7,500

Image formats and re-scaling. Technically, all the images were converted from 2-byte per pixel medical DICOMs format to ordinary 512x512 pixel gray scale representation. Depending on the shape of input images of the CNN architectures being employed, they were re-scaled using B-spline method implemented in the publicly-available software that provides the best possible quality. The re-scaling procedure was applied before the inputting to the CNNs to avoid rough linear on-the-fly interpolation.

Image examples. Examples of pairs of original images of 6 subjects are given in Fig. 1. It is easy to note the age-related changes captured by chest X-Ray images when analyzing the examples column-wise.

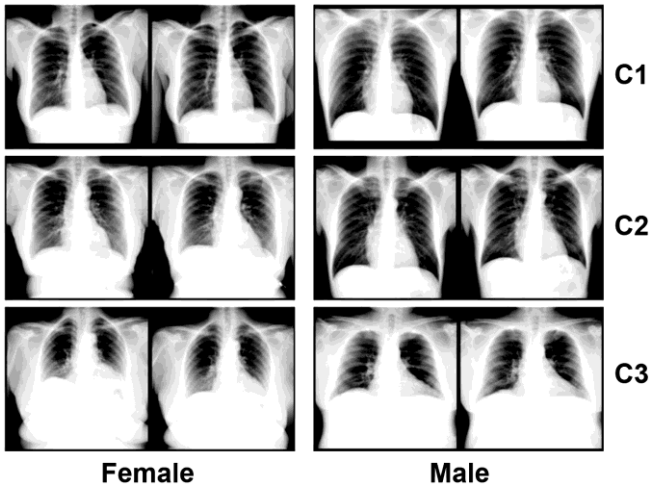

Fig. 1. Examples of original images of Female and Male subjects of Young (C1), Middle (C2), and Mature (C3) age classes

Image sub-sets. The main dataset consisted of 90,000 images was used as a sort of "local repository" of this study which is further sub-divided by subjects' genders and age categories associated with classes C1, C2, and C3. Depending on the specific goal, we created smaller sub-sets by way of appropriate sub-sampling of images from the original C1, C2, and C3 age groups while keeping strongly the necessary age and gender balance. Note that the number of male and female individuals in all the datasets is always represented equally, namely, by 50% of individuals of each gender.

III. METHODS

Image features. It is obvious that different images taken from the same persons are not identical due to various factors. These factors include but not limited to:

- purely technical differences of image acquisition facilities of each hospital including brands of the X-Ray scanners and other engineering environment;

- different poses and possible body movements of persons being scanned; - specific individual condition of subjects such as the volume and the content of their stomach which depends on the amount and type of the food consumed before the examination;

- presence of removable (e.g., gold chains and other kinds of jewelry) as well as non-removable (e.g., implants, cardio stimulators) artefacts;

- changes in person's weight and body proportions happened during the period from the time of previous X-Ray examination till the current one, which could be lasted from only few days and as long as 10-12 years;

- changes of the body and abdominal organs associated with various diseases, traumas, surgeries, and other factors of similar kinds;

- general changes caused by normal ageing, etc.

Thus, searching images by a straightforward pixel-by-pixel comparison is not possible in the problem we are considering here. Instead, it should be performed by way of comparison of image feature vectors representing all the necessary details. The features should capture the key morphological properties of images of different persons while holding certain tolerance to the image variability.

CNN-based feature extraction. In this work, we opted for the use of recent CNNs as the efficient method of feature extraction. Specifically, at this stage of work, we utilized three commonly available CNN architectures including BiT-S R50x1, EfficientNet-B0, and EfficientNet-B0-V2. They produce feature vectors consisted of 2048 and 1280 elements respectively. In addition, the MobileNet-B0-V2 architecture was used for exploratory purposes since it is fast to train and predict.

Some specific properties of the image comparison task considered in this study. In the context of the problem of searching for another medical image of the same person it is worth to keep in mind the points described below.

a) Generally, the image (person) identification is much more complex problem compared to the image classification. This is because in case of identification it is necessary to select one single image among thousands or even millions of similar ones whereas classification typically supposes to categorize every given image to very few classes it appears to be similar to.

b) The complexity of image identification tasks depends on the amount of candidate images in the database. Indeed, let us suppose we have as few as only 2 images in the database. Then, the simple random choice already gives us 50% of success and 2 consequent trials (that is, the Top-2) provide the whole 100%.

c) Under condition of large databases, it is very unlikely the target image would be the most similar in N-dimensional feature space, i.e., it would be the closest to the given query sample by certain metrics.

Thus, the practical approach followed in the computer-assisted identification is typically a two-step procedure. First, a computer-based searching engine is ran to find the Top-N most similar candidates sorted in descending similarity order. Second, a human expert searches the best match manually within the Top-N. Such an approach is widely used in various scenarios of criminal investigations for sample identification, and other real tasks of this sort. We will follow it here too.

d) Note that the manual medical image identification is a very hard and tedious work which takes long time to accomplish. Therefore, in case of large databases the manual identification appears to be not very realistic and even not feasible at all for image datasets larger than few hundreds.

Thus, the appearance of target image in the reasonably short list of Top-N results is highly desirable for any legal user and very displeased for malicious attackers. In this work, the probability of appearance of relevant images in Top-N with the N=80 was the highest which looks like a trade of solution for the formal thresholding. Also, a box-shaped representation of the most similar results supplied to the further visual pair-wise comparison is very convenient for making the final decision.

The key implementation details. The above method was implemented using Python programming language and TensorFlow with Keras libraries. The software was executed on a dedicated server equipped wit 4 GPU of NVIDIA V-100 type with 16 GB of graphics memory each. The statistical analysis procedure that follows was implemented using free R Language and Environment for Statistical Computing [8].

IV. Results

Exploratory analysis of inter-group differences. As it was stated earlier, we are interested not only in the probability of identification of the second image in each pair but also on how such a probability depends on the age and gender factors. Thus, in order to get some idea of how the age image classes are different, we started from a preliminary, exploratory assessment of the difference between the age groups C1, C2, and C3. This was accomplished by way of binary classification of different pair-wise combinations of three classes (there was a reason behind) in form of (C1 vs. C3), (C1 vs. C2), and (C2 vs. C3). Here we used all 3 corresponding datasets as presented in Table 1.

As it was expected, the first pair of (Young vs. Mature) subjects has demonstrated the highest difference with the classification accuracy of 96.2%. Classification of (Young vs. Middle) and (Middle vs. Mature) resulted in lower accuracies of 91.5% and 81.3% respectively.

The presented high classification accuracy achieved in all 3 exploratory experiments confirms that the identification experiments should be performed in a factor-wise manner. These are good news for attackers and bad news for the security staff. Nevertheless, as discussed above, we should keep in mind that the image classification is much less difficult task than the identification one.

Generation of features. All three CNNs were fine-tuned on the image subset called DS-Tune which consisted of 18,000 training images, 3,000 images in each of 6 classes. The final image feature generation step was accomplished using the subset called DS-FTR-gen which contained 24,000 images of 6 classes, 2,000 images in each, 2 scans for each subject. As a result, we got three different feature tables one of which contains 24,000 rows and 2048 columns for BiT-S R50x1 and two other related to two versions of the EfficientNet-B0.

Searching for another image of the same person among all 45,000 people. Features generated by all 3 CNNs were examined and their ability of identification of healthy people by their chest X-Ray images were tested as described above. Results are summarized in form of 3 plots presented in Fig. 2.

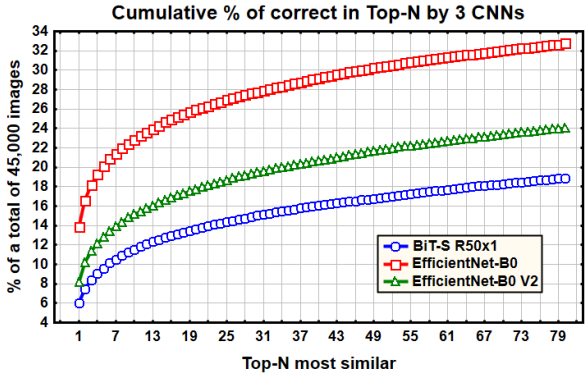

Fig. 2. Percent of correct identification of healthy subjects using features generated by 3 different CNN architectures.

As is easy to see from Fig. 2, features generated by CNN EfficientNet-B0 are significantly better to use in searching for correct image pairs in large X-Ray image databases. This CNN will be used as image feature generator in all the experiments that presented follow.

Identification of Young and Mature (Aged) individuals. Considering the notable differences between the young and mature (Aged) individuals visible on examples presented in Fig.1, we can hypothesize that their individual distinctions should be very different. This guess is confirmed by the data presented in Fig. 3.

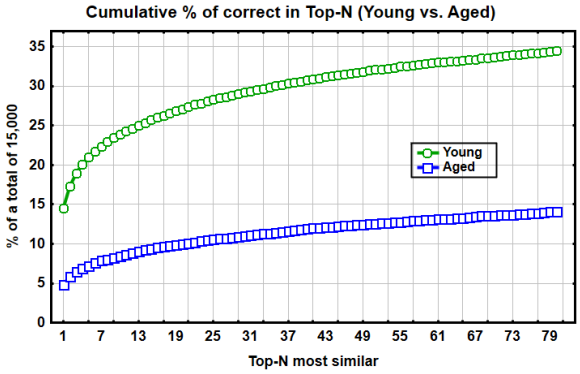

Fig. 3. Percentage of correct identification of Young and Mature (Aged) people of both genders.

Identification of Females and Males. Corresponding results are summarized as plots depicted in Fig. 4.

V. Conclusions

Results obtained with this study allow drawing the following conclusions.

1. The fraction of 45,000 subjects correctly identified by their X-Ray images (appeared on the first position of the top similar) is only about 14% whereas the fraction of correct results in the Top-80 achieves 33%.

2. The chances to be correctly identified are significantly higher for young subjects compared to the mature (aged) ones (14% vs. 5% in Top-1 and 34.5% vs. 14% in Top-80).

3. Percentage of correctly identified females is also higher than in males with the minor difference of 14.5% vs. 13.5% in Top-1 and with a more distinct gap between 36% vs. 30% observed in Top-80.

4. It was found that the CNN EfficientNet-B0 produces better image features distinguishing chest X-Ray images of different people compared to EfficientNet-B0-V2, and BiT-S R50x1. So far, no explanation could be provided for this experimental fact.

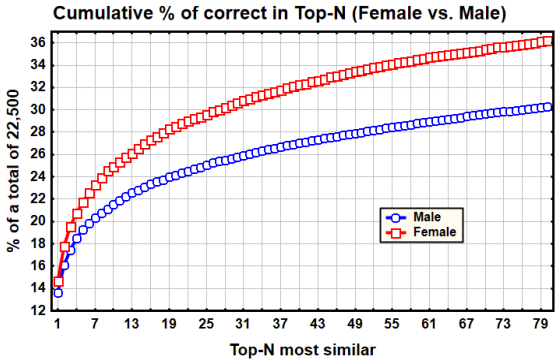

Fig. 4. Percentage of correct identification of Male and Female subjects.

References

[1] V. Liauchuk and V. Kovalev, “Detection of Lung Pathologies Using Deep Convolutional Networks Trained on Large X-ray Chest Screening Database,” Pattern Recognition and Information Processing (PRIP-2019), Minsk, Belarus, 21-23 May, BSUIR, pp. 275-277, 2019.J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[2] V. Kovalev, A. Radzhabov and E. Snezhko, “Automatic detection of pathological changes in chest X-Ray screening images using deep learning methods,” Diagnostic Biomedical Signal and Image Processing Applications, Chapter 8, Elsevier, London, 2023, pp. 155-178.

[3] G. Zhang., B. Liu., T. Zhu, A. Zhu, and W. Zhou , “Visual privacy attacks and defenses in deep learning: a survey,” Artificial Intelligence Review, vol. 55, 2022, pp. 4347–4401. https://doi.org/10.1007/s10462-021-10123-y.

[4] B. Liu, M. Ding, S. Shaham, W. Rahayu, F. Farokhi and Z. Lin, “When Machine Learning Meets Privacy. A Survey and Outlook,” ACM Computing Surveys, vol. 54, issue 2, article 31, March 2022, pp. 1–31.

[5] D. N. Jaidan and L. T. Duong, “Image Features Anonymization for Privacy Aware Machine Learning,” Machine Learning, Optimization, and Data Science, 6th Int Conf, Siena, Italy, July 19–23, 2020, pp. 663–675. https://doi.org/10.1007/978-3-030-64583-0_58.

[6] V. Kovalev, A. Prus and P. Vankevich, “Mining lung shape from X-ray images,” International Conference on Machine Learning and Data Mining (MLDM-2009), Leipzig, Germany, P.Perner (Ed.), LNAI, vol. 5632, Springer, 2009, pp. 554-568.

[7] Web resource, URL: https://image.org.by/, last visited 24.09.2023.

[8] R Core Team (2023). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2023, https://www.R-project.org/.