Identifying cardiomegaly in chest X-rays: A cross-sectional study of evaluation and comparison between different transfer learning methods.

Abstract

Background: Cardiomegaly is a relatively common incidental finding on chest x-rays; if left untreated, it can result in significant complications. Using Artificial Intelligence for diagnosing cardiomegaly, could be beneficial, as this pathology may be under-reported, or overlooked, especially in busy or under-staffed settings.

Purpose: To explore the feasibility of applying four different transfer learning methods to identify the presence of cardiomegaly in chest X-rays and to compare their diagnostic performance using the radiologists' report as the gold standard.

Material and Methods: 2000 chest X-rays were utilised in the current study. 1000 were normal, and 1000 had confirmed cardiomegaly. Of these exams, 80% were used for training and 20% as a holdout test data set. 2048 deep features were extracted using Google's Inception V3 ,VGG16 , VGG19 and SqueezeNet networks. A logistic regression algorithm optimised in regularisation terms was used to classify chest x-rays into those with presence or absence of cardiomegaly.

Results: Diagnostic accuracy is reported by means of sensitivity, specificity, Positive Predictive Value (PPV) and Negative Predictive Value (NPV), with the VGG19 network providing the best values of sensitivity (84%), specificity (83%), PPV (83%), NPV (84%) and overall accuracy (84,5%). The other networks presented sensitivity between 64.1%-82%, specificity between 77.1%-81.1%, PPV between 74%-81.4%, NPV between 68%-82% and overall accuracy between 71%-81.3%.

Conclusion: Deep learning using transfer learning methods based on VGG19 network can be used for automatic detection of cardiomegaly on chest X-ray images. However, further validation and training of each method is required before application to clinical cases.

Keywords: Artificial Intelligence, Deep Learning, Transfer Learning, Cardiomegaly, Validation.

Introduction

Background:

Cardiomegaly is a condition defined by the enlargement of the heart, while it is thought to be associated with other diseases such as hypertension, coronary artery disease, kidney disease and cardiomyopathy (1,2). Cardiomegaly can be present in various genetic and acquired cardiomyopathies. The prevalence of hypertrophic cardiomyopathy (HCM) ranges between 1/250 to 1/500 in adults, while the prevalence of dilated cardiomyopathy (DCM) seems to be two-fold. In children, HCM is uncommon, while DCM is more likely to occur in the first year of life (3). Cardiomegaly can originate from various abnormalities, such as hypertension, thyroid disorders, heart valve or coronary artery disease. Early detection is important to ensure a modified diet or exercise regime is in place to reduce further complications.

Plain chest x-rays have been established as a standard imaging method to detect cardiomegaly, using the Cardio-Thoracic Ratio (CTR) as the parameter to evaluate, quantify and diagnose this pathology (4). CTR is calculated as the ratio of the transverse diameter of the heart to the maximum diameter of the internal thoracic cavity. **Specifically, the maximum diameter of the internal thoracic cavity is calculated by measuring the diameter from the medial border of the ribs (4). Usually, this measurement is performed on the level of the dome of the right hemidiaphragm. The transverse diameter of the heart is calculated after measurement of the right and left most borders of the heart, horizontally (fig.1).** A CTR of 50% is thought to be the upper normal limit, although 45% has been also suggested (4,5). Although novel imaging techniques are nowadays available, plain chest x-rays are still the most cost-effective and accessible method to detect enlargement of the cardiac silhouette, while the specificity of this method when using CTR as the measurement parameter has proved to be high (84.5%) (6). Therefore, early and accurate detection of cardiomegaly on plain chest x-rays is considered to be very important and may also help reduce the costs associated with more expensive imaging methods. In this study, Transfer Learning (TL) methods were employed and evaluated to facilitate the detection of cardiomegaly on plain chest x-rays. **The purpose of this study is to explore the feasibility of applying four different transfer learning methods, to identify the presence of cardiomegaly in chest X-rays and to compare their diagnostic performance using the radiologists' report as the gold standard.**

Introduction to Transfer learning networks:

The recent technological advancements have brought Artificial Intelligence (AI) at the forefront of Medicine and Radiology.

The most widely used branch of AI in medicine is Machine Learning (ML), a statistical process to generate knowledge by training models with data, and fitting models to data (7). With ML we can now provide systems with the ability to undertake complex tasks with high accuracy without even being explicitly programmed by identifying patterns in streams of input data (8). Deep Learning (DL), a fast-growing branch of ML since 2010s, attempts the abstraction of features directly from raw data, using multi-layered Deep Neural Networks (DNNs) (fig.2) (9).

DL can be categorised into supervised, semi-supervised and unsupervised (fig.2) (10). DL is thought to be a powerful tool for use within healthcare, however, it is commonly accepted that it is generally limited by issues such as low quality, low volume and high sparsity of input data, which may limit the performance of DL methods (11). DL learning methods include the use of Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), two different types of learning methods. Many CNN-based models are used for classification of free-text reports, of mammograms, assessment of skeletal bone age, organ segmentation during MRI-guided radiotherapies (12-15). However, one of the most important problem of machine learning which we encounter after we have trained our model is when the model matches the training data almost perfectly but performs poorly in validation of new data. This is widely known as overfitting. Overfitting could be avoided by down sampled operations on the input maps, which reduces the dimension of the data.

Transfer learning refers to a method of reusing pre-trained models or knowledge for solving another related task avoiding using training from scratch. As it is widely known, training a model from scratch requires a vast amount of data in order to accomplish a high prediction level. Using transfer learning enables us to use the weights of an already trained model such ImageNet which consists of a dataset of about 1.2 million images for training, 50,000 for validation and 100,000 for testing, belonging to 1000 categories, without its last fully connected layer as a feature extractor (16,17).

The advantage of TL is that the weights are already pre-trained, instead of using training from scratch (10). Therefore, these models are fine-tuned from pre-trained models, instead of using random Gaussian distributions or learning from scratch, without its final layer as a fixed feature extractor (fig.3). This fine-tuning process enables to fine-tune all the kernels by means of back-propagation. TL is a method approach commonly used within DL applications, and it has proved to be very effective, especially with small training datasets (18). DL is very efficient in learning discriminative features and learn directly from the raw data (11). It can effectively extract learning patterns and data from large, complex data. DL-based algorithms are widely used in radiology, offering the ability of object recognition, classification, localization and segmentation in medical images (9). When training a small network, usually a pre-trained large dataset is employed. In this study, ImageNet was used as a pre-trained model, offering the advantages of containing 14 million images with 1000 classes, while it is also publicly accessible (19,20). Transfer learning strategies depend on various factors, but the most important ones are the size of the new dataset, and its similarity to the original one.

More detailed information on common strategies used in transfer learning can be found in supplementary material.

Material and Methods

a) Study design:

This is a retrospective quantitative research study, as it makes predictions based on the analysis of already acquired data (21). In this retrospective study, the pre-trained model approach was used to apply transfer learning (22). Ethics approval was not required as open-source data was used, readily available. No identifiable patient information was recovered. The STROBE cross-sectional reporting guidelines were used to write this article (23).

b) Databases used:

Large-scale image classification dataset:

ImageNet was chosen in this study as the large-scale object recognition dataset, on which the compared networks had been pre-trained. Its feasibility as a large-scale image recognition dataset is well-established within the literature (24). ImageNet provides more than 14 million of images with 1000 classes, more than 20,000 categories and it has been widely used during TL applications within medical imaging (20, 25).

Classification method:

In this current study off-the-shelf transfer learning was used removing the last fully connected layer of the pre-trained models and replacing it with a logistic regression classifier optimised with regularisation.

Logistic regression:

Logistic regression is a supervised learning method, which is usually preferred for binary classification problems (two class values) as in this study and is used to model the probability of a certain class (normality vs abnormality, in this case cardiomegaly). In terms of mathematics, a binary logistic model has a dependent variable with two possible values 0 or 1. To shorten the coefficients of the outcome of the classification a regularisation was used. As input variables were changed, the model's prediction changes a lot. Regression in terms of regularisation reduces the variance of the model and the prediction rate is higher (26).

c) Pre-trained convolutional neural networks tested:

Four different TL methods were evaluated and compared in this study. For this reason, four different pre-trained DL image recognition models with different amount of convolution layers were used. All of the models used in this study were trained on ImageNet database. These models are Google's Inception V3, VGG16, VGG19 and SqueezeNet (27). The Inception V3 network is a widely used network, offering accuracy of greater than 78.1% on the ImageNet dataset, widely used within medical imaging, with very good results (28,29). The VGG19 is a CNN comprising of 19 deep layers, it is trained on more than one million images on the ImageNet dataset, and it has shown promising results when applied to CNN-based methods within Radiology (30,31). The third deep neural network used in this study was SqueezeNet, also trained on ImageNet database, consisting of 18 deep layers, and having the ability to classify images into 1000 categories (32). Lastly, the VGG16 neural network was used in this study. This is a 16-layer network which has been widely used in similar studies, offering the advantage of an optimal performance during the Large-Scale Visual Recognition Challenge (24).

d) Data where transfer-learning methods were tested:

A total of 2000 plain chest x-rays were retrospectively collected from the Picture Archival and Communications System (PACS). These images were categorized into 2 categories, based on the presence or absence of cardiomegaly. Images with poor quality or no clear anatomic presentation of the organ-target were excluded from the set. Among these x-rays, 1000 were normal (regarding the size of the cardiac silhouette) and 1000 of them depicted a confirmed cardiomegaly. 1600 plain chest x-rays (80%) were used for training, while the remaining 400 (20%) images were used as test set. 2048 deep features were extracted from each plain chest x-ray image (fig. 4). A logistic regression algorithm optimized in regularisation terms was used to classify chest x-rays into presence or absence of cardiomegaly (normal or abnormal).

The selected plain chest x-rays were extracted from the viewer and converted into 384x384 pixel-sized images with a JPEG format. No further processing was applied to the images with regards to size and shape. The following figure (fig.5) demonstrates an example of our data with a normal chest radiograph and a radiograph with cardiomegaly. **It must be noted that there were cases in which the used networks**

failed to correctly identify the images as normal or abnormal, and such cases are demonstrated in the following figure (fig.6).

e) Data analysis and evaluation of diagnostic performance indices:

Sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV) were all calculated as part of evaluating overall diagnostic accuracy of each method (33,34). This method was preferred for evaluation of the accuracy instead of ROC analysis, as it is more suitable for balanced data like the ones used for this study.

Results

Using the remaining 400 plain chest x-rays of the test set, the performance of Google's Inception V3, VGG16, VGG19 and SqueezeNet neural networks was evaluated.

With regards to the sensitivity (the true positive rate) of the evaluated networks, Google's Inception V3 network achieved a sensitivity of 64.1%, while the VGG16 network had a sensitivity rate of 81%. Similarly, the sensitivity of SqueezeNet was reported at 82%. However, the sensitivity that the VGG19 network achieved was the highest among the tested networks, achieving a sensitivity rate of 84%. Consequently, the sensitivity of the tested networks achieved similar rates, between 81-84%, with the exception of Google's Inception V3, which yielded a significantly lower sensitivity rate. Figure 7 and table 1 depict the overall distributions of the tested networks regarding different accuracy indicators.

The measurements of the specificity of each of the tested networks, showed that the networks achieved a more similar specificity compared to the sensitivity, as there is a greater uniformity among them. However, the lowest specificity rate was noted again for Google's Inception V3 network, achieving a rate of 77.1%. Similarly, the highest specificity rate was achieved by VGG19 network (83%), while VGG16 and SqueezeNet achieved 81.1% and 80% respectively.

The PPVs of each utilized network were assessed, and that indicated that again the VGG19 network achieved the highest PPV among all networks, giving a PPV of 83%. On the contrary, the lowest PPV was reported for Google's Inception V3 network (74%), while the VGG16 and SqueezeNet achieved a PPV of 81,4% and 80% respectively.

NPV is a measure to assess the degree of the true negative results of a test, meaning that a negative prediction of the test is yielded, and the true condition is also negative. The highest NPV was achieved by the VGG19 network (84%), while Google's Inception V3 was again the network with the lowest rate (68%). Similarly, VGG16 and SqueezeNet achieved similar results regarding NPV (81% and 82% respectively), giving the conclusion that the overall performance related to NPV was similar to PPV.

Finally, the overall accuracy (OA) of the tested networks was evaluated. Overall accuracy is the weighted average of a test's sensitivity and specificity (33). The evaluation of the OA of the tested neural networks showed that the VGG19 network achieved the highest OA among all the networks, giving an OA of 84,5%. On the contrary, Google's Inception V3 network achieved the lowest OA among the four tested networks (71%). Following the results of sensitivity and specificity, the VGG16 and SqueezeNet networks achieved similar rates of OA, reporting an OA of 81,3% and 81% respectively.

Discussion

There is currently a wealth of academic literature underlining the feasibility and efficacy of TL approaches using CNNs for image recognition and classification (20,25,35).

Previous studies have used other deep learning methods, with variable results in detecting cardiomegaly but also using a smaller sample size (36-39). Another study has tested 3 transfer learning algorithms including Inception 3 but found out much lower accuracy levels than the ones we describe here (40). The results of our study clearly indicate that a TL approach using CNNs can be reliably used for automatic detection of cardiomegaly on plain chest x-rays. Moreover, after evaluating the performance of four different networks on the same data set, this study revealed that the VGG19 neural network has the highest diagnostic performance among these networks. The superiority of the VGG19 network has been already supported, with Shaha and Pawar concluding that a fine-tuned VGG19 network achieves an overall higher performance compared to VGG16 (41).

Our study reinforces published literature (37,42,43), indicating that automatic detection of cardiomegaly is now feasible when using pre-trained neural networks. These methods are very promising and already revolutionise feature exaction from medical images.

However, there are still some limitations:

- This study uses a pre-trained network, and this may differ from actual medical images in many different ways.
- We need not have access to prior medical images or other clinical data of these patients to be able to understand the clinical significance, correlations or

importance of these findings; this would be better suited to a prospective study.

- It also uses a relatively small dataset. Within the literature, it is a consensus that when applying CNN-based approaches to medical imaging, limited datasets can be a great drawback, and growing a well-annotated dataset is believed to be as crucial as developing new algorithms (17). However, using TL, small datasets can be used, taking a neural network, which has been pre-trained on a large dataset and adapt its knowledge to the specific task without over-fitting (44).
- Such models must be able to become generalisable to unseen data. To achieve this, over-fitting must be minimised. Over-fitting occurs when a model memorises the noise instead of the signal on the image. Therefore, this over-fitted model will not perform well on a new dataset (20). Increasing the size of the dataset will result in decreased over-fitting.
- This method is able to differentiate between normal appearance of the cardiac silhouette and severe cardiomegaly. Further work must be done to ensure that discrimination of marginal cases of cardiomegaly would not be missed by the proposed network.

Although the results of the current study are very promising for the timely evaluation of cardiomegaly on plain chest x-rays using TL methods, future work with larger datasets and use of prospective studies with new clinical imaging data are needed in order to establish the optimal methods to be used. Moreover, either fine-tuning transfer learning or using learning from scratch, would offer the ability to use larger data sets, compare the results and suggest the optimal methods for the diagnosis of cardiomegaly.

In conclusion, Transfer Learning approaches using pre-trained CNNs have been well established for image classification, recognition and segmentation. The results of our study confirm the efficacy of these methods for automatic detection of cardiomegaly on plain chest x-rays and conclude that the pre-trained network VGG19 has superior performance compared to three other networks. Further research is needed using larger prospectively collected datasets for clinical validation of the model.

References

1. Alghamdi SS, Abdelaziz I, Albadri M, et al. Study of cardiomegaly using chest x-ray. J Radiat Res Appl Sc 2020;13:460-467.
2. Paulman PM, Paulman AA, Harrison JD, et al. Taylor's differential diagnosis manual: symptoms and signs in the time-limited encounter. Philadelphia: Wolters Kluwer, 2013.
3. McKenna WJ, Maron BJ, Thiene G. Classification, epidemiology, and global burden of cardiomyopathies. Circ Res 2017;121:722-730.
4. Brakohiapa EKK, Botwe BO, Sarkodie BD, et al. Radiographic determination of cardiomegaly using cardiothoracic ratio and transverse cardiac diameter: can one size fit all? Part one. Pan Afr Med J 2017;27.
5. Mensah YB, Mensah K, Asiamah S, et al. Establishing the cardiothoracic ratio using chest radiographs in an indigenous Ghanaian population: a simple tool for cardiomegaly screening. Ghana Med J 2015;49:159-164.
6. Biharas Monfared A, Agha Farajollah S, Sabour F, et al. Comparison of radiological findings of chest x-ray with echocardiography in determination of the heart size. Iran Red Crescent Med J 2015;17:e18242.

7. Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. Future Healthc J 2019;6:94-98.
8. Sidey-Gibbons JAM, Sidey-Gibbons CJ. Machine learning in medicine: a practical introduction. BMC Med Res Methodol 2019;19:64.
9. Cao C, Lin F, Tan H, et al. Deep learning and its applications in biomedicine. Genom Proteom Bioinf 2018;16:17-32.
10. Alom Z, Taha TM, Yakopcic C, et al. A state-of-the-art survey on deep learning theory and architectures. Electronics 2019;8: 292.
11. Chen Q, Hu S, Long P, et al. A transfer learning approach for malignant prostate lesion detection on multiparametric MRI. Technol Cancer Res Treat 2019;18.
12. Chen MC, Bell RL, Yang L, et al. Deep learning to classify radiology free-text reports. Radiology 2018;286:845-852.
13. Jadoon MM, Zhang Q, Haq IV, et al. Three-class mammogram classification based on descriptive CNN features. Biomed Res Int 2017;3640901.
14. Spampinato C, Palazzo S, Giordano D, et al. Deep learning for automated skeletal bone age assessment in X-ray images. Med image anal 2017;36:41-51.
15. Fu Y, Mazur TR, Wu X, et al. A novel MRI segmentation method using CNN-based correction network for MRI-guided adaptive radiotherapy. Med Phys 2018;45:5129-5137.
16. Zeiler MD, Fergus R. Visualizing and understanding convolutional networks. In: Fleet D, Pajdla T, Schiele B, et al eds. Computer Vision – ECCV 2014. Lecture Notes in Computer Science. Switzerland: Springer International Publishing, 2014:818-833.

17. Shin H-C, Roth HR, Gao M, et al. Deep convolutional neural networks for computer-aided detection: cnn architectures, dataset characteristics and transfer learning. IEEE Trans Med Imaging 2016;35:1285-1298.
18. Maqsood M, Nazir F, Khan U, et al. Transfer learning assisted classification and detection of Alzheimer's disease stages using 3D MRI scans. Sensors (Basel) 2019;19:2645.
19. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Commun ACM 2017;60:84-90.
20. Yamashita R, Nishio M, Do RKG, et al. Convolutional neural networks: an overview and application in radiology. Insights Imaging 2018;9:611-629.
21. Gerrish K, Lathlean J. The research process in nursing. 7th ed. United Kingdom: Willey-Blackwell, 2015.
22. Yim J, Joo D, Bae J, et al. A gift from knowledge distillation: fast optimization, network minimization and transfer learning. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017 July 21-26; IEEE, 2017:7130-7138.
23. von Elm E, Altman DG, Egger M, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. J Clin Epidemiol 2008;61:344-349.
24. Russakovsky O, Deng J, Su H, et al. ImageNet large scale visual recognition challenge. Int J Comput Vis 2015;115:211-252.
25. Lakhani P, Sundarem B. Deep learning at chest radiography: automated classification of pulmonary tuberculosis using convolutional neural networks. Radiology 2017;284:574-582.

26. Miller M. The basics: logistic regression and regularization. Available at: https://towardsdatascience.com/the-basics-logistic-regression-and-regularization-828b0d2d206c (accessed 10 July 2020).

27. Kurama V. A Review of Popular Deep Learning Architectures: resNet, InceptionV3, and SqueezeNet. Available at: https://blog.paperspace.com/popular-deep-learning-architectures-resnet-inceptionv3-squeezenet/ (accessed 1st July 2020).

28. Song J, Chai YJ, Masuoka H, et al. Ultrasound image analysis using deep learning algorithm for the diagnosis of thyroid nodules. Medicine 2019;98:e15133.

29. Kang G, Liu K, Hou B, et al. 3D multi-view convolutional neural networks for lung nodule classification. PLoS One 2017;12:e0188290.

30. Kim HG, Lee KM, Kim EJ, et al. Improvement diagnostic accuracy of sinusitis recognition in paranasal sinus x-ray using multiple deep learning models. Quant imaging med surg 2019;9:942-951.

31. Ahammed Muneer KV, Rajendran VR, Paul Joseph K. Glioma tumor grade identification using artificial intelligent techniques. J Med Syst 2019;43:113.

32. Kaggle Competitions. Available at: https://www.kaggle.com/ (accessed 1st July 2020).

33. Trevethan R. Sensitivity, specificity, and predictive values: foundations, pliabilities, and pitfalls in research and practice. Front Public Health 2017;5:307.

34. Habib A, Alalyani M, Hussain I, et al. Brief review on sensitivity, specificity and predictivities. IOSR-JDMS 2015;14:64-68.

35. Han D, Liu Q, Fan W. A new image classification method using CNN transfer learning and web data augmentation. Expert Syst Appl 2018;95:43-56.
36. Rajpukar P, Irvin J, Ball RL, et al. Deep learning for chest radiograph diagnosis: a retrospective comparison of the CheXneXt algorithm to practicing radiologists. PLoS Med 2018;15:e1002686.
37. Que Q, Tang Z, Wang R, et al. CardioXNet: automated detection for cardiomegaly based on deep learning. Proceedings of the 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC); 2018 July 18-21; Honolulu. IEEE, 2018:612-615.
38. Seah JCY, Tang JSN, Kitchen A, et al. Chest radiographs in congestive heart failure: visualizing neural network learning. Radiology 2019;290:514-522.
39. Cicero M, Bilbily A, Colak E, et al. Training and validating a deep convolutional neural network for computer-aided detection and classification of abnormalities on frontal chest radiographs. Invest Radiol 2017;52:281-287.
40. Zhu S, Zhang X, Zhang R. Identifying cardiomegaly in ChestX-ray8 using transfer learning. Stud Health Technol Inform 2019;264:482-486.
41. Shaha M, Pawar M. Transfer Learning for Image Classification. Proceedings of the Second International Conference on Electronics, Communication and Aerospace Technology (ICECA); 2018 March 29-31; Coimbatore. IEEE, 2018:656-660.
42. Candemir S, Rajaraman S, Thoma G, et al. Deep learning for grading cardiomegaly severity in chest x-rays: an investigation. Proceedings of the IEE Life Sciences Conference (LSC); 2018 Oct 28-31; Montreal. IEEE, 2018:109-113.

43. Mohare S, Dawani H, Rathi V, et al. Detection of cardiomegaly from chest X-rays using Otsu algorithm and convolutional neural network. Proceedings of the IEEE 5th International Conference for Convergence in Technology (I2CT); 2019 March 29-31; Bombay. IEEE, 2019:1-5.

44. Vakli P, Deak-Meszlenyi RJ, Hermann P, et al. Transfer learning improves resting-state functional connectivity pattern analysis using convolutional neural networks. Gigascience 2018;7:giy130.

Table 1. Summary of overall performance of the tested networks

	Sensitivity	**Specificity**	**PPV**	**NPV**	**OA**
Inception V3	64,1%	77,1%	74%	68%	71%
VGG-16	81%	81,1%	81,4%	81%	81,3%
VGG-19	84%	83%	83%	84%	84,5%
SqueezeNet	82%	80%	80%	82%	81%

Figure 1. The standard measurements for calculating the CTR measurement.

Figure 2. The hierarchy of Artificial Intelligence and its main branches.

Figure 3. Training procedure with Deep Learning and Transfer Learning.

Figure 4. Graphical depiction of the study's pipeline.

Figure 5. Normal chest radiograph (Fig. 5A) and cardiomegaly (Fig.5B) (32).

Figure 6. Cases in which the networks failed to correctly identify cardiomegaly. A. Normal result instead of cardiomegaly, due to suboptimal image quality. B, C. Results of cardiomegaly despite the normal size of the heart silhouette, due to suboptimal image quality (B) and catheter insertion (C).

Figure 7. Summary of overall performance of the tested networks.