2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

Deep Learning and Binary Relevance Classification of Multiple Diseases using Chest X-Ray images*

Marc-André Blais and Moulay A. Akhloufi, *Senior Member IEEE*

***Abstract*— Disease detection using chest X-ray (CXR) images is one of the most popular radiology methods to diagnose diseases through a visual inspection of abnormal symptoms in the lung region. A wide variety of diseases such as pneumonia, heart failure and lung cancer can be detected using CXRs. Although CXRs can show the symptoms of a variety of diseases, detecting and manually classifying those diseases can be difficult and time-consuming adding to clinicians' work burden. Research shows that nearly 90% of mistakes made in a lung cancer diagnosis involved chest radiography. A variety of algorithms and computer-assisted diagnosis tools (CAD) were proposed to assist radiologists in the interpretation of medical images to reduce diagnosis errors. In this work, we propose a deep learning approach to screen multiple diseases using more than 220,000 images from the CheXpert dataset. The proposed binary relevance approach using Deep Convolutional Neural Networks (CNNs) achieves high performance results and outperforms past published work in this area.**

***Clinical relevance*— This application can be used to support physicians ans speed-up the diagnosis work. The proposed CAD can increase the confidence in the diagnosis or suggest a second opinion. The CAD can also be used in emergency situations when a radiologist is not available immediately.**

I. Introduction

Chest X-ray (CXR) is widely used in detecting a variety of diseases affecting the chest area. This technology can help doctors detect a variety of diseases, such as a pneumonia, pulmonary edema, heart failure, lesions, lung cancer, tuberculosis, sarcoidosis, and pleural effusion. Furthermore, the possibility of screening a disease (E.g. cancer) using a CXR can augment the survival rate of patients as shown in various studies [1], [2]. This makes the CXR highly useful given its availability in almost all clinics compared to other methods.

However, when analyzing the results of the chest images, a variety of components may complicate the analysis of the images. The lack of specialized personnel to analyze the images or fatigue can lead to errors The inconsistency in diagnosis by radiologists can also be a major issue since the interpretation of a CXR may differ from one specialist to another.

A CAD system can be used to help reduce the burden on radiologists while reducing the possibility of errors. A variety of CAD systems have already proven their usefulness on an extensive array of diseases [3], [4], [5], [6]. Furthermore, Convolutional Neural Networks (CNN) have shown great promise in the medical field when it comes to disease classification [7], [8], [9]. However, most of current research focuses on a single label classification while fro CXR we are interested in multi-label classification. Multi-label classification is a situation where an image may have one or more diseases present. This increases the complexity of the problem since the algorithm must be able to detect multiple diseases even if they overlap. Previous research using the CheXpert dataset [10] explored the idea of multi-label classification. In [11], the authors used a directed acyclic graph (DAG) approach with deep CNN (DCNN) to learn dependencies between classes.

*This research was enabled in part by support provided by the New Brunswick Health Research Foundation (NBHRF), Calcul Québec (calculquebec.ca) and Compute Canada (www.computecanada.ca)

M.A. Blais and M.A. Akhloufi are with the Perception, Robotics, and Intelligent Machines Research Group (PRIME), Dept of Computer Science, Université de Moncton, Moncton, NB, Canada {emb9357, moulay.akhloufi}@umoncton.ca

II. Related Work

Convolutional Neural Networks (CNN) have shown their performance and proven their efficacy in detecting and classifying diseases in a vast array of imaging modalities. zhang *et al.* [12] were able to detect benign and malignant breast tumors in shear-wave elastography with an accuracy of 93.4% and an Area Under Curve (AUC) of 0.947 using a deep learning approach. Similarly, Huynh *et al.* [7] used deep learning to detect mammography tumors. Using transfer learning, they trained the AlexNet architecture and were able to achieve an AUC of 0.86. Mohsen *et al.* [13], proposed the use of deep learning to classify brain tumors into four categories (normal, glioblastoma, sarcoma and metastatic bronchogenic carcinoma tumor). They were able to achieve a mean classification accuracy of 96.97% and a mean AUC of 0.984. Finally, Han *et al.* [14] tested the ability of deep learning to classify and detect 12 skin diseases. Using ResNet-152 they achieved an average AUC of 0.91 on the Asian Test Dataset and an AUC of 0.89 on the Edinburgh Dataset. Rajpurkar *et al.* [15] used ChestX-ray14 dataset [16] to detect 14 pneumonia types. They used binary relevance and DenseNet-121. Although this method reduces the complexity of the training, it does not directly account for the relation between classes. This method achieved an average AUC of 0.8411 and an F1-score of 0.435 which is higher than the radiologists (0.387). In a similar approach, Narin *et al.* [17] used a variety of DCNNs to detect Covid-19. Using a small dataset of 50 images, they were able to achieve an F1-score of 1 on both the InceptionV3 and ResNet-50 architecture. In another work with 192 images [18], Chetoui *et al.* obtained an AUC of 0.973, a specificity of 0.966 and a sensitivity of 0.951 using ResNet-50.

A. DCNNs on CheXpert

The dataset used in this work was published and accompanied by a deep learning approach to classify the diseases [10].

This paper also explored 5 methods to deal with the uncertainty label that can be present in the training subset. The U-Ignore adjustment consists of ignoring the label that are marked as Uncertain (-1) while calculating the loss during the training phase. This method can be viewed as the safer option since it doesn't speculate the unknown labels like the other three methods. The second and third method is to set the uncertainty labels as either 0 (U-zero method) or 1 (U-One method). However, as one can expect, this method could mislabel a large proportion of the classes which would then misrepresent the results. The fourth method called U-SelfTrained has a similar concept to a semi-supervised model using the U-ignore method. This method consists of first training a model using the U-ignore approach until convergence. Then the labels that have an uncertainty are predicted by the model who was previously trained. Similar to the second and third method, some labels may be misclassified which would greatly impact the final results. The fifth method named U-MultiClass consist of dealing with the uncertainty labels as its own class. This method completely removes the problem of creating miss-classified labels while keeping all the images plus 33% more classes. This addition of classes not only augments the complexity of the model but also creates a redundancy in labels.

All five methods were trained using a DenseNet-121 [19] and Adam optimizer [20] on images resized to 320x320 pixels. A variety of other models were also used such as Alexnet [21] and Resnet-50 [22] nevertheless the DenseNet performed the best. However, this paper couldn't come to a conclusion regarding the best method for handling the uncertainty labels. Some methods have a significant augmentation for a particular disease while for a different disease the difference is minimal. This paper achieved a maximum AUC of 0.858 for Atelectasis (U-one), 0.854 for Cardiomegaly (U-MultiClass), 0.939 for Consolidation (U-SelfTrained), 0.935 for Edema (U-SelfTrained), 0.936 for Pleural Effusion (U-MultiClass) which gives an average AUC of 0.9054.

B. Directed Acyclic Graph

Pham *et al.* [11] proposed a hierarchical like method of training using what is known as a Directed Acyclic Graph (DAG). The proposed approach has the current best overall performance for the CheXpert dataset.

DAG is a method to represent the dependencies between diseases such as the presence of an enlarged cardiomediastinum in cardiomegalies. This explicit implementation of dependencies and the effect of it were not taken into account in previous research. Using a DAG allows the human to attribute a known correlation between labels which reduces the amount of information a DCNN must learn. The first step of this method is to learn the dependencies between the parent disease and its leaf using a CNN. This is done by only using positive parents to train a model able to classify its direct child, meaning a two level child leaf can not be classified using this method. The second step consists of freezing all but the last layer of the CNN and then to train this layer with the full dataset. The prediction of a class is thus the conditional probability of the label being positive and the parent of said label to be also positive. A Bayes rule is then used to find the unconditional probability of a class by multiplying the conditional probabilities of all parents of a child.

This method however does not directly remove the uncertainty class from the images which is still the biggest challenge of this dataset. To solve this, a method called label smoothing regularization (LSR) combined with previous methods proposed in [10]. This method consists of using either U-Zero or U-One from [10] and replacing the absolute 0 and 1 by a random float close to the value of the respective method. I.e. when using the U-Zero method then a value would be closer to 0 rather than 1 (e.g. 0.05) while the U-Zero would be closer to 1 (e.g. 0.95). This has the purpose of reducing the impact of the uncertainty label which in turns reduces the impact of mislabeled data. However this method does not eliminate the possibility of having mislabeled data, it only reduces it.

The images of size 224x224 were first normalized using the mean and standard deviation from the ImageNet dataset [23]. A variety of models (DenseNet-121,169,201 [19], Inception-ResNet-v2 [24] and Xception [25]) were then combined to create an ensemble learning model. This method achieved an AUC of 0.909 for Atelectasis, 0.910 for Cardiomegaly, 0.957 for Consolidation, 0.958 for Edema, 0.964 for Pleural Effusion with a mean AUC of 0.940. These results show the possibility of using deep learning to detect and classify a variety of diseases present on CXRs.

III. Dataset

CheXpert dataset [10] was collected by the Stanford Hospital between 2002 and 2017 from a variety of patients leading to over 200,000 images. With 14 diseases, this dataset had images which may contain multiple positive labels on a single image. The labels are as follows: support device (105,831), lung opacity (92,669) , pleural effusion (75,696), edema (48,905), atelectasis (29,333), cardiomegaly (23,002), pneumothorax (17,313), no finding (16,627), consolidation (12,730), enlarged cardiom. (9,020), fracture (7,270), lung lesion (6,856), pneumonia (4,576) and pleural other (2,441). The dataset consisted of 220,000 training images which had either a frontal or side view of the chest. Some example images are given in figure 1. A validation set of 200 images was also included, this set of images only contained positive and negative labels.

The ground-truth value of the images consists of four possible labels which are **u**, **0**, **1** or **-1**. The labeling of the images were done by an automatic labeler using radiologists reports. The positive (1) and negative labels (0) means that a disease was present or not. The **u** label means that all mentions of the disease were negative but one mention was uncertain. The uncertainty label (-1) represents either the

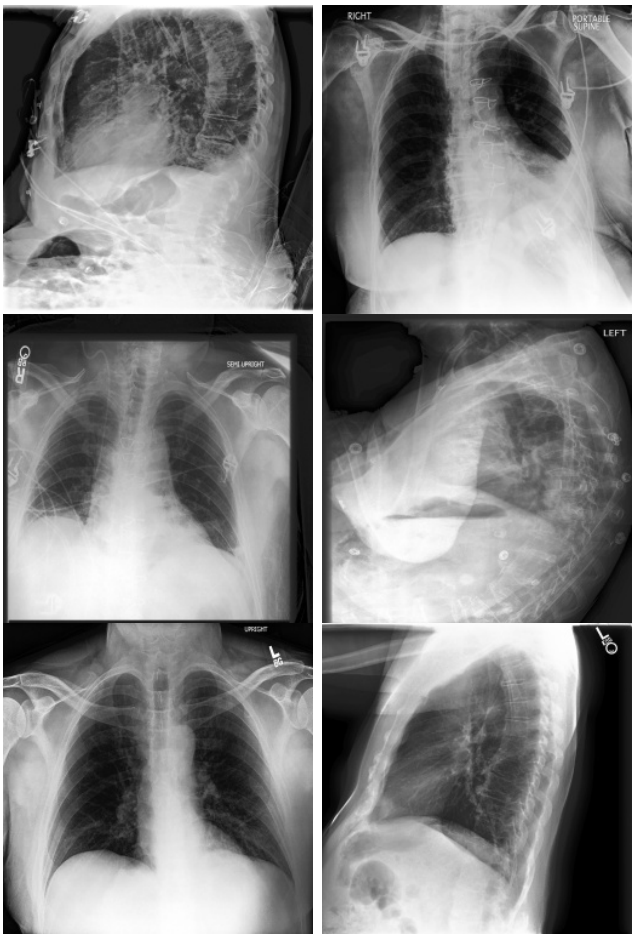

Fig. 1: Example images from CheXpert dataset, from left to right, up to down: Atelectasis, consolidation, cardiomegaly, pleural effusion, edema, and no finding.

uncertainty of a radiologist or the ambiguity in a report. These two labels represent a challenge since they couldn't be attributed to a specific class.

A. Dataset Modification

The images in both training and validation sets were resized to 512x512 while reducing the grayscale image from three channels to one channel. The resizing and color space modification were the only modifications done on this dataset to not affect the validity of the images.

For the labels we decided to set all **u** labels as negative since statistically they have a higher chance of being negative. For the uncertainty label (-1) we decided to opt for the safest approach and ignore those images during the training phase. This allowed us to be more confident with our results since there was no ambiguity in the used labels. Since we are using binary relevance, removing the images with an uncertain label did not greatly impact the size of the disease specific subset. We used 15,000 images from the training subset as our validation set.

The previous papers using this dataset only referred to five diseases (edema, consolidation, cardiomegaly, pleural effusion and atelectasis) to validate their approach. This was due to both the important number of these disease images and the imbalance within the given validation dataset. To compare with previous research, we decided to also use those five diseases as an indicator of the overall performance of our approach.

IV. Proposed Approach

A. Binary Relevance

Previous research on CheXpert dataset used what is commonly known as multi-label classification. This method of classification is when a singular model is trained to predict N classes where N is the number of labels. Although this method has its advantages such as limiting the number of models to one, this is however also its main disadvantage. Due to the need of predicting all the classes out of one model, the complexity increases drastically and the performance decreases. In this work, we use a binary relevance approach. Binary relevance can be viewed as having N unique models, each predicting a distinctive class for N labels. Binary relevance allows us to more efficiently deal with the uncertainty label by ignoring it. During the training and validation phase of our models, we ignore the images with an uncertainty label for its disease specific model. Considering the size of our dataset, ignoring a portion of the images is not a problem and assures the certitude of the positive and negative labels used. We used the Area Under the Curve (AUC) as a performance metric. This metric is used due to its ability of giving a bigger importance to a minority class such as our positive labels in an imbalanced dataset.

B. Models

The following architectures were adapted to our problem: DenseNet-121,169,201 [19], InceptionResnetV2 [24], InceptionV3 [26], MobileNet, MobileNetV2 [27], ResNet101, ResNet101V2, ResNet152, ResNet152V2, ResNet50, ResNet50V2 [22], VGG16, VGG19 [28] and Xception [25].

These models were both trained with Adam [20] and SGD optimizers [29]. We tested both training the full architecture using the available training set and also used transfer learning with pretrained layers on the ImageNet dataset [21].

V. Results

The Adam optimizer achieved its best results combined with the pretrained Xception architecture with an average AUC of **0.9587**. The results for the individual disease detection using the Adam optimizer are as follows: an AUC of 0.9390 for Atelectasis, 0.9667 for Cardiomegaly, 0.9469 for Consolidation, 0.9647 for Edema and 0.9766 for Pleural Effusion. Unlike Adam optimizer, the SGD optimizer did not achieve an overall best result with one model (different architectures performed differently on each disease). The SGD optimizer achieved an AUC of 0.9535 for Atelectasis (ResNet152-pretrained), 0.9721 for Cardiomegaly (Res152V2-pretrained), 0.9575 for Consolidation (Xception-pretrained), 0.9729 for Edema (Res50V2-pretrained) and 0.9822 for Pleural Effusion (ResNet152-pretrained). These

results give an average AUC of **0.9676** compared to 0.9587 from the Adam trained models.

When comparing with state-of-the-art work, we can see that we obtain a higher AUC for all the five measured diseases. The previous highest mean AUC achieved was from [11] with an average AUC of 0.940 while our best average AUC is **0.9676**. This increase of 0.0276 (2.76%) using a simple architecture shows the improved performance given by the deep binary relevance approach proposed in this work.

In addition we tested the models on all 14 diseases using the Xception network with Adam optimizer and achieved a mean AUC of 0.949. The AUC result for each class are as follows: support device (0.96855), lung opacity (0.92172) , pleural effusion (0.9748), edema (0.96436), atelectasis (0.93883), cardiomegaly (0.96885), pneumothorax (0.96354), no finding (0.95858), consolidation (0.94346), enlarged cardiom. (0.92827), fracture (0.9357), lung lesion (0.93618), pneumonia (0.9406) and pleural other (0.94478).

VI. Conclusion

We propose a simple but effective approach for the detection and classification of multiple diseases using CXR images. Using the CheXpert dataset, we developed a deep Convolutional Neural Network techniques which could effectively classify the CXR images. Our approach achieved higher results on all the five diseases used in past works. The proposed deep binary relevance approach explains the achieved performance.

The developed techniques can be used to build a CAD system to help radiologists and physicians and speed up the diagnosis. Furthermore, this method could easily be converted for the detection of other pulmonary diseases. Future work includes developing an ensemble approach with the developed models, optimizing and creating new algorithms, testing more datasets, and adapting the approaches to other imaging modalities and other diseases.

References

[1] G. Gavelli and E. Giampalma, "Sensitivity and specificity of chest x-ray screening for lung cancer," *Cancer*, vol. 89, no. S11, pp. 2453–2456, 2000.

[2] G. M. Strauss, R. E. Gleason, and D. J. Sugarbaker, "Chest x-ray screening improves outcome in lung cancer: a reappraisal of randomized trials on lung cancer screening," *Chest*, vol. 107, no. 6, pp. 270S–279S, 1995.

[3] K. Doi, "Computer-aided diagnosis in medical imaging: historical review, current status and future potential," *Computerized medical imaging and graphics*, vol. 31, no. 4-5, pp. 198–211, 2007.

[4] B. Van Ginneken, B. T. H. Romeny, and M. A. Viergever, "Computer-aided diagnosis in chest radiography: a survey," *IEEE Transactions on medical imaging*, vol. 20, no. 12, pp. 1228–1241, 2001.

[5] F. De Dombal, D. Leaper, J. R. Staniland, A. McCann, and J. C. Horrocks, "Computer-aided diagnosis of acute abdominal pain," *Br Med J*, vol. 2, no. 5804, pp. 9–13, 1972.

[6] M. P. Sampat, M. K. Markey, A. C. Bovik, *et al.*, "Computer-aided detection and diagnosis in mammography," *Handbook of image and video processing*, vol. 2, no. 1, pp. 1195–1217, 2005.

[7] B. Q. Huynh, H. Li, and M. L. Giger, "Digital mammographic tumor classification using transfer learning from deep convolutional neural networks," *Journal of Medical Imaging*, vol. 3, no. 3, p. 034501, 2016.

[8] M. Anthimopoulos, S. Christodoulidis, L. Ebner, A. Christe, and S. Mougiakakou, "Lung pattern classification for interstitial lung diseases using a deep convolutional neural network," *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1207–1216, 2016.

[9] Q. Li, W. Cai, X. Wang, Y. Zhou, D. D. Feng, and M. Chen, "Medical image classification with convolutional neural network," in *2014 13th international conference on control automation robotics & vision (ICARCV)*, pp. 844–848, IEEE, 2014.

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

[11] H. H. Pham, T. T. Le, D. Q. Tran, D. T. Ngo, and H. Q. Nguyen, "Interpreting chest x-rays via cnns that exploit disease dependencies and uncertainty labels," *arXiv preprint arXiv:1911.06475*, 2019.

[12] Q. Zhang, Y. Xiao, W. Dai, J. Suo, C. Wang, J. Shi, and H. Zheng, "Deep learning based classification of breast tumors with shear-wave elastography," *Ultrasonics*, vol. 72, pp. 150–157, 2016.

[13] H. Mohsen, E.-S. A. El-Dahshan, E.-S. M. El-Horbaty, and A.-B. M. Salem, "Classification using deep learning neural networks for brain tumors," *Future Computing and Informatics Journal*, vol. 3, no. 1, pp. 68–71, 2018.

[14] S. S. Han, M. S. Kim, W. Lim, G. H. Park, I. Park, and S. E. Chang, "Classification of the clinical images for benign and malignant cutaneous tumors using a deep learning algorithm," *Journal of Investigative Dermatology*, vol. 138, no. 7, pp. 1529–1538, 2018.

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

[16] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. Summers, "Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *IEEE CVPR*, 2017.

[17] A. Narin, C. Kaya, and Z. Pamuk, "Automatic detection of coronavirus disease (covid-19) using x-ray images and deep convolutional neural networks," *arXiv preprint arXiv:2003.10849*, 2020.

[18] M. Chetoui, A. Traoré, and M. A. Akhloufi, "Deep learning for covid-19 detection on chest x-ray and CT scan," in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, IEEE, 2020. Poster.

[19] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.

[20] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[21] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, pp. 1097–1105, 2012.

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[23] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *CVPR09*, 2009.

[24] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Thirty-first AAAI conference on artificial intelligence*, 2017.

[25] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1251–1258, 2017.

[26] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

[27] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[28] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[29] H. Robbins and S. Monro, "A stochastic approximation method," *The annals of mathematical statistics*, pp. 400–407, 1951.