Attention-driven Spatial Transformer Network for Abnormality Detection in Chest X-Ray Images

Joana Rocha
INESC-TEC and Faculty of Engineering, University of Porto
R. Dr. Roberto Frias s/n, 4200-465,
Porto, Portugal.
0000-0002-4856-138X
joana.m.rocha@inesctec.pt

Sofia Cardoso Pereira
INESC-TEC and Faculty of Engineering, University of Porto
R. Dr. Roberto Frias s/n, 4200-465,
Porto, Portugal.
0000-0001-6754-6495

João Pedrosa
INESC-TEC and Faculty of Engineering, University of Porto
R. Dr. Roberto Frias s/n, 4200-465,
Porto, Portugal.
0000-0002-7588-8927

Aurélio Campilho
INESC-TEC and Faculty of Engineering, University of Porto
R. Dr. Roberto Frias s/n, 4200-465,
Porto, Portugal.
0000-0002-5317-6275

Ana Maria Mendonça
INESC-TEC and Faculty of Engineering, University of Porto
R. Dr. Roberto Frias s/n, 4200-465,
Porto, Portugal.
0000-0002-4319-738X

2022 IEEE 35th International Symposium on Computer-Based Medical Systems (CBMS) | 978-1-6654-6770-4/22/$31.00 ©2022 IEEE | DOI: 10.1109/CBMS55023.2022.00051

***Abstract*—Backed by more powerful computational resources and optimized training routines, deep learning models have attained unprecedented performance in extracting information from chest X-ray data. Preceding other tasks, an automated abnormality detection stage can be useful to prioritize certain exams and enable a more efficient clinical workflow. However, the presence of image artifacts such as lettering often generates a harmful bias in the classifier, leading to an increase of false positive results. Consequently, healthcare would benefit from a system that selects the thoracic region of interest prior to deciding whether an image is possibly pathologic. The current work tackles this binary classification exercise using an attention-driven and spatially unsupervised Spatial Transformer Network (STN). The results indicate that the STN achieves similar results to using YOLO-cropped images, with fewer computational expenses and without the need for localization labels. More specifically, the system is able to distinguish between normal and abnormal CheXpert images with a mean AUC of 84.22%.**

***Index Terms*—binary classification, deep learning, module, object detection, radiography, thorax**

I. Introduction

Among the most popular medical imaging exams, the Chest X-Ray (CXR) is frequently requested by healthcare professionals to assess the presence of thoracic diseases, due to its low-cost and non-invasive nature. Nevertheless, a thorough analysis of CXR images is time-consuming and their interpretation may be dubious even for expert radiologists [1]. For this reason, the incorporation of computer-aided diagnosis systems in the hospitals is an attractive solution to provide a second opinion, and promote greater efficiency in the interpretation of these exams. Following the advances in computational capabilities and the increasing availability of medical data sets, Deep Learning (DL) based systems can provide a great preliminary diagnostic tools to reduce the physicians' workload. In particular, considering that cardiothoracic and pulmonary abnormalities are one of the leading causes of morbidity and mortality worldwide [2], a CXR-based abnormality detection system may help clinicians to prioritize more urgent abnormal exams. Hence, this work focuses on a system for the detection of general thoracic abnormalities, using the binary image-level labels "normal" or "abnormal".

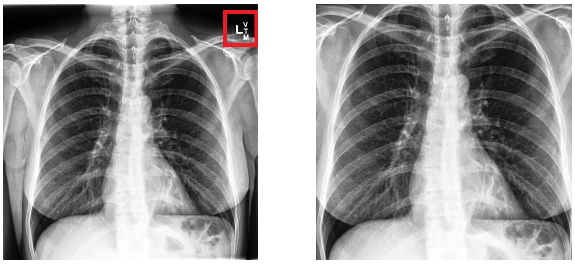
While DL is mainly known for its embedded feature engineering, inherently selecting and combining attributes for a more efficient learning process, the resulting models are also subject to data-related biases [3]. For instance, the letters present in a CXR scan may be wrongfully interpreted as pathologic features (Figure 1a in red) [4]. Such characteristic is a major setback for their broad adoption in a clinical setting, whose domain is grounded in the ability to exploit causal relationships. This way, researchers often resort to a preprocessing stage in which they crop the images' Region of Interest (ROI) before feeding them to a classifier, selecting only the portion of the image that contains the thorax. Object detection models are often used for this purpose, as exemplified in Figure 1b. By selecting the ROI and removing image artifacts, this approach may reduce the number of false positive predictions, but at the same time it also implies training a completely separate model with bounding box ground truths before training the main classifier, thus creating a significant extra computational cost.

This work was funded by the ERDF - European Regional Development Fund, through the *Programa Operacional Regional do Norte* (NORTE 2020) and by National Funds through the FCT - Portuguese Foundation for Science and Technology, I.P. within the scope of the CMU Portugal Program (*NORTE-01-0247-FEDER-045905*) and *LA/P/0063/2020*. The work of J. Rocha was supported by the FCT grant contract *2020.06595.BD*. The work of S. Pereira was supported by the FCT grant contract *2020.10169.BD*.

2372-9198/22/$31.00 ©2022 IEEE
DOI 10.1109/CBMS55023.2022.00051

Spatial Transformer Networks (STNs) were originally introduced in [5] and refer to any Convolutional Neural Network (CNN) that includes a Spatial Transformer module. This module can be trained end-to-end with a classifier to transform the original input images in a way that benefits the classification task. In sum, the present work focuses on replacing the typical object detection model [6], [7] for a more efficient workflow, using an attention-driven STN to simultaneously select the ROI and classify the images, and thus suppressing the need for two independent models. The main contributions of this paper are summarized below:

1) A single STN whose attention-driven module and classifier are trained end-to-end to crop and classify the CXR scans, without any modifications to the standard loss function.
2) A reduced computational cost in comparison to the typical object detection models, which require localization ground truth annotations and larger architectures and training times.



(a) Original image with artifacts marked in red.

(b) ROI selected by a thorax detection model.

Fig. 1: Example of an image before and after cropping.

II. Related Work

Published work in this field has typically favored pathology classification rather than abnormality detection; yet, such detection task can have a high impact when it comes to building a triage system for the analysis of CXR images. In general, standard off-the-shelf CNN-based methods are frequently applied to the data in question, establishing a comparison between well-known architectures [8], [9]. In alternative, other state-of-the-art publications opt for autoencoder-based one-class learning approaches in order to infer on normality patterns and distinguish the abnormal scans [10], [11].

Only two relevant papers have recently implemented STNs in CXR scans to achieve an invariant canonical pose through affine transformations (rotation, shear, scaling, and translation). More specifically, the first one focuses on multi-label pathology classification using spatial annotations, and proposes a loss function that minimizes the difference between the transformed scans and a canonical CXR reference [12]. The second addresses the binary abnormality classification using multi-modal data and image-level labels [13]. Both of these publications employ the NIH ChestX-Ray14 dataset [14].

TABLE I: Label distribution of the CheXpert subset.

Findings	Frequency	Binary Setting
Atelectasis	6,440	22,146 abnormal images
Cardiomegaly	4,044	
Consolidation	4,312	
Edema	2,547	
Enlarged Cardiomediastinum	3,297	
Fracture	1,540	
Lung Lesion	2,482	
Lung Opacity	10,794	
Pleural Effusion	9,880	
Pleural Other	1,776	
Pneumonia	3,469	
Pneumothorax	2,223	
No Finding	5,508	5,508 normal images

In other domains, STNs have been used mostly for object localization [15] and scene recognition [16], [17], benefiting from the attention provided by the module, but resorting to overall complex architectures with multiple branches to detect more than a single object.

III. Methodology

A. Dataset and Preprocessing

The dataset employed in this work comprises a total of 27,654 postero-anterior CXR scans selected from the CheXpert database. The scans were provided by the Stanford ML Group and the corresponding annotations include image-level labels of common thoracic diseases [18]. More specifically, the frequency of each pathologic finding in this subset is discriminated in Table I, yielding 5,508 normal images (class "0") and 22,146 abnormal images (class "1"). The conversion of the original annotations into binary ones was done following a *U-Ones* approach, in which uncertain labels are considered positive, and all images were resized to 512×512 pixels. Note that the presence of support devices was disregarded as is not indicative of any abnormality, and that an image was considered normal only if none of the other findings are mentioned, independently of the presence of these devices.

B. Experimental Setup

Three experimentation settings were established. The first two are considered baseline experiments and are used as control for comparison purposes with the third and main setting, which is the proposed STN. In all cases, the results were achieved by splitting the dataset across five folds for cross-validation, preserving the original pathology proportion and without patient overlap. Three folds were used for training, one for validation, and one for testing. Additionally, the training data was augmented to generate more samples through soft random affine transformations (i.e. 5 degrees of rotation range and 3 degrees of shear range). To tackle the considerable class imbalance, a weighted random sampler was implemented when loading the training set. Figure 2 shows the general workflow for the three experimentation settings, which are further described below.

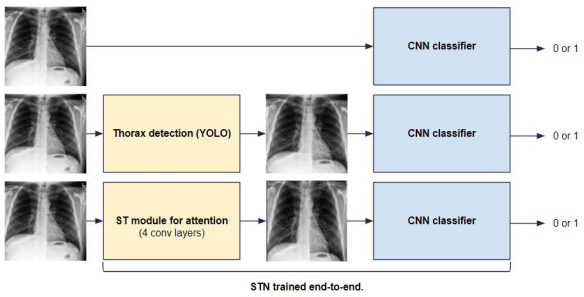


Fig. 2: Workflow of the baseline, baseline with thorax detection, and STN.

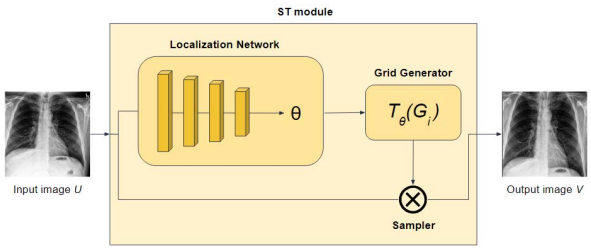


Fig. 3: Schematic of the ST module.

1) Baseline: The baseline setting corresponds to the implementation of a standard binary classifier, a widely-used VGG16 architecture, whose last fully-connected layer was modified to consider a single node. This layer is followed by a sigmoid activation function to yield the corresponding predictions. Other architectures were initially considered, but ultimately the VGG16 proved to have enough layers for this specific learning task. Therefore, the model was trained using the binary cross entropy loss and Adam optimizer, with an initial learning rate of 10^{-5} and a total of 20 epochs (due to the fast convergence of the network). This classifier is common to all the following experimentation settings, preserving the same optimization routine.

2) Baseline with Thorax Detection: This section focuses on refining the initial baseline workflow by restricting the image regions being analysed by the model - more specifically, by keeping only the thorax portion of the scan and eliminating any structures outside this ROI. Note that the aim here is to prevent the influence of information represented outside the thorax region, such as left/right lettering markers or irrelevant anatomical features (e.g. the patient's neck and shoulders).

The thorax detection model was based on a YOLOv5 and trained with 956 CXRs from publicly available lung segmentation databases. More specifically, the ground-truth was obtained by drawing a bounding box around the manual lung field segmentation masks provided in the JSRT, Montgomery, and Shenzhen datasets [19], [20]. The YOLOv5 network was initialized using weights pre-trained on the COCO dataset [21], and training was performed using a stochastic gradient descent optimizer with an initial learning rate of 10^{-2} for a duration of 150 epochs.

Once trained, the detection model was applied to the CheXpert data to yield bounding boxes of the thorax region. Low confidence predictions (i.e. probability inferior to 0.90) were discarded to keep only the most relevant cropped images as reference. The largest dimension of each remaining bounding box was used to crop a square centered in the middle of the box, without distorting the proportions of the image. The images were then resized back to 512×512 pixels.

3) Spatial Transformer Network: The STN is composed of the ST module, followed by the classifier previously outlined in Section III-B1. Hence, the present section will focus on the description of the module and its main components: a localization network, a grid generator, and a sampler, represented in Figure 3.

Firstly, an input image U is fed to the localization network that predicts the theta parameters (θ). These parameters encode the affine transformations to be applied to the image, and originally encompassed operations of scaling, translation, rotation, and shearing. However, the last two operations are not necessary for the purpose of attention, and so it is proposed to reduce the complexity of the transformations by predicting only four θ parameters instead of six, corresponding to vertical and horizontal scaling (s) and translation (t) in Equation 1. Here, a simple CNN was considered for the architecture of the localization network, composed of four convolutional blocks (convolutional layer with ReLU activation followed by max pooling), a dense layer with ReLU activation, and a final dense layer to predict the four θ variables.

The domain knowledge indicates that a larger zoom in the vertical axis is needed rather than in the horizontal axis. For this reason, non-isotropic scaling was considered to better frame the ROI, particularly since the original scans already have shorter margins on the left and right sides of the thorax, in comparison to the top and bottom margins. This way, the parameters were initialized as follows: $s_x = 1$, $s_y = 0.1$, $t_x = 0$, and $t_y = 0$. Note that the original publication proposes to initialize all parameters with an identity matrix, meaning that the model will implement no transformations in the first epoch, and then progress towards optimal theta values during training [5]. However, in this case the s_y component can be initialized with a low positive value (0.1), meaning that early transformations start with a large vertical zoom and progress towards the optimal s_y prediction in the following epochs. This enables a faster convergence, as 0.1 is closer than 1 to that optimal s_y value.

$$\theta = \begin{bmatrix} \theta_1 & \theta_2 & \theta_3 \\ \theta_4 & \theta_5 & \theta_6 \end{bmatrix} = \begin{bmatrix} s_x & 0 & t_x \\ 0 & s_y & t_y \end{bmatrix} \quad (1)$$

The grid generator will then obtain a sampling grid G_i with target coordinates (x_i^t, y_i^t), based on the predicted θ transformation matrix. Finally, the sampler uses bilinear interpolation

in the set of sampling points $T_\theta(G_i)$ to produce the final output *V*. This process is depicted in Equation 2, that determines the source image coordinates (x_i^s, y_i^s) to transform.

$$\begin{pmatrix} x_i^s \\ y_i^s \end{pmatrix} = T_\theta(G_i) = \begin{bmatrix} s_x & 0 & t_x \\ 0 & s_y & t_y \end{bmatrix} \times \begin{pmatrix} x_i^t \\ y_i^t \\ 1 \end{pmatrix} \quad (2)$$

In short, the module yields a transformed image with the same size as the original input, which is passed along to the VGG16 for classification, in a way that both the module and the classifier are optimized together and trained end-to-end.

IV. Results and Discussion

This section analyses the results of the three scenarios previously described, in order to infer if it is possible for the STN to achieve similar results to using YOLO-cropped images, with less computational expense and without the need for localization labels. To do so, the analysis focuses on two main aspects: the performance of the classifier in these settings, and the results of the ROI selection using the YOLO model and the ST module.

A. Classifier Performance

The VGG16 classifiers were evaluated using the following standard metrics: Area under the ROC Curve (AUC), Precision-Recall Area under the Curve (PR-AUC), and precision. For a more accurate comparison between the three scenarios, a high recall operating point of 95% was defined. The high recall suits the medical domain premise, in which a false negative prediction is more severe than a false positive one. Table II presents the average results across five different test sets, including the rate of True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN) predictions.

Overall, the STN classifier outperformed both baseline settings in all established metrics: the baseline with thorax detection offers a 0.77% increase in AUC in comparison to the single classifier, while the STN offers an increase of 1.01%. Additionally and as expected, both the YOLO model and the ST module contributed to reduce the FP rate and increase the TN rate, by eliminating misleading image artifacts. The latter was able to reduce the FP rate in 0.97% and boost the TN rate in 0.96%.

No significant changes were detected in terms of TPs and FNs. While the classifier is not able to discriminate specific types of abnormal findings, the top-three pathologies that most contribute to the TP rate are edema, pneumonia, and cardiomegaly/lung opacity (similar score), and the top-three pathologies that most contribute to the FN rate are fractures, enlarged cardiomediastinum, and lung lesions.

Direct comparisons with the state-of-the-art are challenging, due to the lack of publications addressing the same binary classification setting with an STN-based approach and/or the same data. Nevertheless, there is one publication which employs an STN to distinguish between normal and abnormal scans [13]. This publication reports a precision increase of 1% between their baseline and the STN, but does not disclose AUC scores. Regarding abnormality detection in general, all papers cited in Section II employ the NIH ChestX-ray14 dataset, in opposition to the more recent CheXpert one. The justification for choosing CheXpert over ChestX-ray14 in this study is grounded in [22]. Both cases generate most of the annotations through natural language processing labelers, but their overall consistency and reliability is indisputably different. For instance, the author argues that ChestX-ray14 lacks patient diversity and proper labeling structure and quality, with error rates that range from 30% to 90% in the various findings. The error rates in CheXpert are approximately between 5% and 15%. The author also concludes that CheXpert is larger and overall better for DL implementation, as its labeler was more thoroughly evaluated and proved to produce "labels that accurately reflect the reports".

TABLE II: Mean and standard deviation results after cross-validation.

%	**Baseline**	**Baseline with Thorax Detection**	**Proposed STN**
AUC	83.21±0.69	83.98±0.75	**84.22**±0.70
PR-AUC	94.78±0.29	95.12±0.28	**95.23**±0.29
Precision	84.33±0.39	85.00±0.42	**85.25**±0.49
TP Rate	76.06±0.04	76.05±0.01	**76.08**±0.01
FP Rate	14.14±0.50	13.43±0.44	**13.17**±0.42
TN Rate	5.78±0.96	6.49±0.44	**6.74**±0.42
FN Rate	4.02±0.05	4.03±0.02	**4.01**±0.02

B. ROI Selection

Both the YOLO detection model and the ST module are employed for attention, forcing the classifier to infer on the image regions most relevant for the diagnosis. The main difference between these two approaches is the fact that the first is trained independently from the classifier using spatial annotations (bounding boxes locating the thorax), and the second is optimized simultaneously with the classifier, using exclusively the image-level binary labels. Several examples of both approaches are presented in Figure 4.

Regarding the YOLO model, an average precision of 99.84% at an IoU>0.5 was obtained in the validation set. This detection network is able to perfectly delimit the left and right boundaries of the thorax, but on the other hand, by preserving the original proportion of the images, the vertical cropping is not as precise, thus not being able to eliminate certain artifacts in Figures 4a,d,e.

As mentioned in Section III-B3, the STN considers non-isotropic scaling, meaning the original proportions of the image may or may not be preserved. In this study, the classifier benefited from a larger vertical scaling than a horizontal one, increasing the detail in that direction. In fact, the horizontal delimitation of the module is not as precise as in YOLO, but rather focusing its attention on the center and lower regions of the thorax, as showed in Figures 4b,c. In other words, the

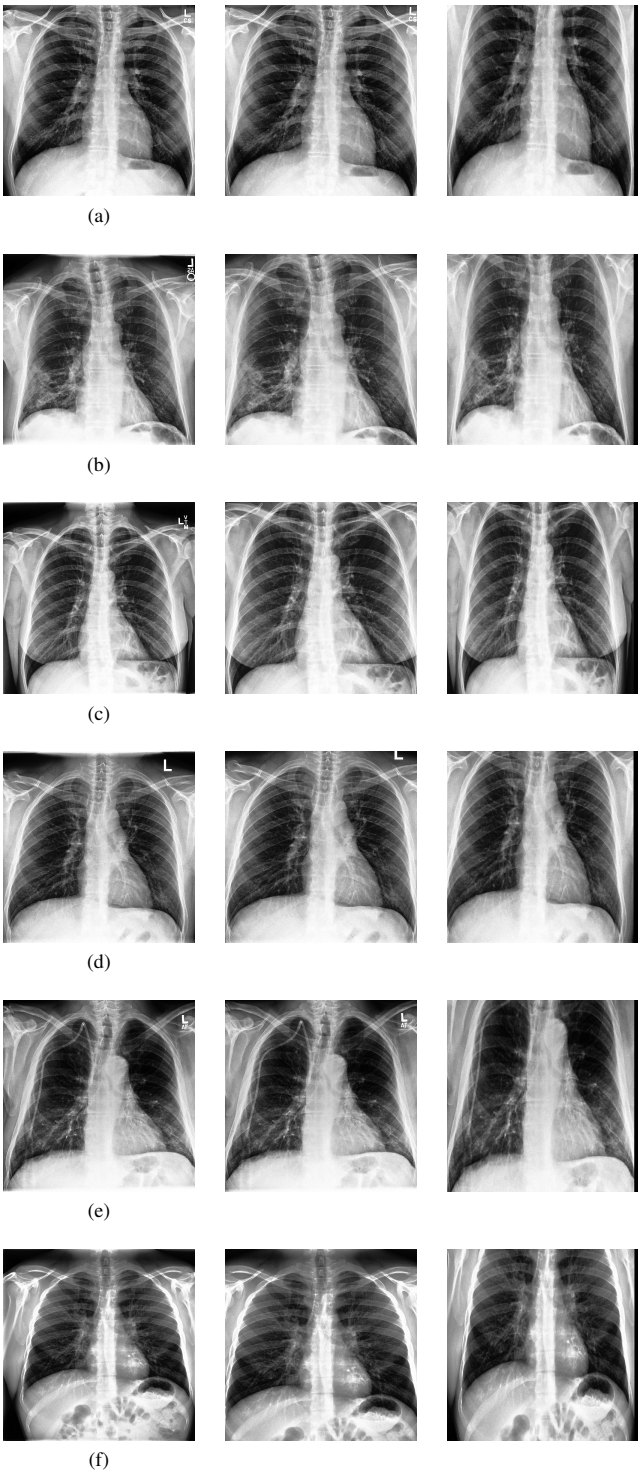


Fig. 4: Examples per row of original, cropped (YOLO), and transformed (STN) images, respectively.

STN is able to successfully delimit the bottom of the lungs, but also tends to slightly crop the lung apex. As the network is trained in a spatially unsupervised manner to select the ROI, it may infer that there are no statistically relevant features in the upper part of the lungs. This is more severe when the lower abdomen representation is noisier, drawing the model's attention to that bottom region (Figure 4f).

The proposed method differs from the state-of-the-art in the sense that no changes were made to the loss function, and no multi-modal data was employed. In fact, the main distinction from the methods proposed in [12], [13] is the restriction in allowed transformations, which enabled the attention-driven component of the work, while using a much simpler architecture for the module. The STN results presented in Table II indicate an improvement in precision of 0.92% versus the standard baseline, similar to the 1% improvement disclosed in [13]. Note that different datasets were employed, so a direct comparison cannot be established.

V. Conclusions

The successful classification of thoracic abnormalities in CXR scans can be extremely useful for patient management and follow-up, when implemented as a triage step to prioritize more urgent cases. Moreover, being able to do so in an automated fashion can bring crucial advantages in a clinical context, in which resources are often scarce. The current work aims to contribute to the detection of abnormal images through an automated deep learning based binary classification model, to help the specialists distinguishing between normal and abnormal instances.

Consequently, the developed architecture faces three major challenges: the inherent complexity associated with the anatomical structures, the often noisy, blurred and/or low contrasted nature of medical images, and the presence of image artifacts that mislead the model into mostly FP predictions. To tackle these difficulties, an attention-driven approach is considered as a means of restricting the learning process exclusively to the area of interest to the diagnosis. In CXR data analysis, this becomes particularly relevant to ensure that the classifier is in fact making predictions based on the correct anatomical region, in this case the thorax. Note that this work targets only image artifacts found outside the ROI, and does not consider the artifacts that may be present within that region, such as arrows or other annotations.

The proposed approach can be described as a CNN, composed of an attention-driven ST module and a classifier, which is able to select the thoracic ROI, transform the input images accordingly, and then detect several indiscriminate types of abnormalities in a binary manner. This is a more efficient alternative to using an object detection model as an *a priori* processing stage, promoting the use of a single network to perform all tasks end-to-end. Furthermore, the present publication demonstrates that the STN is able to replace the complex YOLO architecture with a four convolutional layer module to achieve similar cropping results without the need for two independent models. Generally speaking, by restricting

the transformation matrix to only translation and non-isotropic scaling, the model dynamically scales and aligns the images to maximize the classifier's performance. The non-isotropic nature of the theta parameters allow for a more detailed scaling of the input images, thus eliminating more artifacts than the traditional approach. In sum, the proposed approach is beneficial in terms of required computational power and training time, without the need for spatial annotations.

While the proposed model succeeds at identifying the lower limit of the lungs and cropping the ROI accordingly, it is not as precise at identifying the upper limit. This aspect may be regarded in future work, as measures can be taken to prevent the loss of information in the lung apex, in order to preserve the totality of the thorax for further analysis. Such step would ensure no minor abnormality is neglected by the model during inference, and may be implemented by adding further restrictions to the scaling components.

REFERENCES

[1] N. J. Shaw, M. Hendry, and O. B. Eden, "Inter-Observer Variation in Interpretation of Chest X-Rays," *Scottish Medical Journal*, vol. 35, no. 5, pp. 140–141, Oct. 1990. [Online]. Available: http://journals.sagepub.com/doi/10.1177/003693309003500505

[2] H. Wang, M. Naghavi, C. Allen, R. M. Barber, and Z. Bhutta et. al, "Global, regional,and national life expectancy, all-cause mortality, and cause-specific mortality for 249 causes of death,1980–2015: a systematic analysis for the Global Burden of Disease Study 2015," *388*, pp. 1459–1544, 2016. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0140673616310121

[3] M. A. Gianfrancesco, S. Tamang, J. Yazdany, and G. Schmajuk, "Potential Biases in Machine Learning Algorithms Using Electronic Health Record Data," *JAMA Internal Medicine*, vol. 178, no. 11, pp. 1544–1547, Nov. 2018. [Online]. Available: https://doi.org/10.1001/jamainternmed.2018.3763

[4] A. J. DeGrave, J. D. Janizek, and S.-I. Lee, "AI for radiographic COVID-19 detection selects shortcuts over signal," *Nature Machine Intelligence*, vol. 3, no. 7, pp. 610–619, Jul. 2021. [Online]. Available: http://www.nature.com/articles/s42256-021-00338-7

[5] M. Jaderberg, K. Simonyan, A. Zisserman, and k. kavukcuoglu, "Spatial Transformer Networks," in *Advances in Neural Information Processing Systems*, vol. 28. Curran Associates, Inc., 2015. [Online]. Available: https://proceedings.neurips.cc/paper/2015/hash/33ceb07bf4eeb3da587e268d663aba1a-Abstract.html

[6] E. Sogancioglu, E. Çallı, B. van Ginneken, K. G. van Leeuwen, and K. Murphy, "Deep Learning for Chest X-ray Analysis: A Survey," *Medical Image Analysis*, vol. 72, p. 102125, Aug. 2021, arXiv: 2103.08700. [Online]. Available: http://arxiv.org/abs/2103.08700

[7] A. Karacı, "VGGCOV19-NET: automatic detection of COVID-19 cases from X-ray images using modified VGG19 CNN architecture and YOLO algorithm," *Neural Computing and Applications*, vol. 34, no. 10, pp. 8253–8274, May 2022. [Online]. Available: https://link.springer.com/10.1007/s00521-022-06918-x

[8] Y.-X. Tang, Y.-B. Tang, Y. Peng, K. Yan, M. Bagheri, B. A. Redd, C. J. Brandon, Z. Lu, M. Han, J. Xiao, and R. M. Summers, "Automated abnormality classification of chest radiographs using deep convolutional neural networks," *npj Digital Medicine*, vol. 3, no. 1, p. 70, Dec. 2020. [Online]. Available: http://www.nature.com/articles/s41746-020-0273-z

[9] E. J. Yates, L. C. Yates, and H. Harvey, "Machine learning "red dot": open-source, cloud, deep convolutional neural networks in chest radiograph binary normality classification," *Clinical Radiology*, vol. 73, no. 9, pp. 827–831, Sep. 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S000992601830206X

[10] Y. Mao, F.-F. Xue, R. Wang, J. Zhang, W.-S. Zheng, and H. Liu, "Abnormality Detection in Chest X-Ray Images Using Uncertainty Prediction Autoencoders," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, A. L. Martel, P. Abolmaesumi, D. Stoyanov, D. Mateus, M. A. Zuluaga, S. K. Zhou, D. Racoceanu, and L. Joskowicz, Eds. Cham: Springer International Publishing, 2020, pp. 529–538.

[11] N. Shvetsova, B. Bakker, I. Fedulova, H. Schulz, and D. V. Dylov, "Anomaly Detection in Medical Imaging With Deep Perceptual Autoencoders," *IEEE Access*, vol. 9, pp. 118 571–118 583, 2021, conference Name: IEEE Access.

[12] J. Liu, G. Zhao, Y. Fei, M. Zhang, Y. Wang, and Y. Yu, "Align, Attend and Locate: Chest X-Ray Diagnosis via Contrast Induced Attention Network With Limited Supervision," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2019.

[13] S. Bharati, P. Podder, and M. R. H. Mondal, "Hybrid deep learning for detecting lung diseases from X-ray images," *Informatics in Medicine Unlocked*, vol. 20, p. 100391, Jan. 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352914820300290

[14] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "ChestX-ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017.

[15] A. Meethal, M. Pedersoli, S. Belharbi, and E. Granger, "Convolutional STN for Weakly Supervised Object Localization," in *2020 25th International Conference on Pattern Recognition (ICPR)*, Jan. 2021, pp. 10 157–10 164, iSSN: 1051-4651.

[16] S. Guo, L. Liu, W. Wang, S. Lao, and L. Wang, "An attention model based on spatial transformers for scene recognition," in *2016 23rd International Conference on Pattern Recognition (ICPR)*, Dec. 2016, pp. 3757–3762.

[17] D. Liu, Y. Wang, and J. Kato, "Supervised Spatial Transformer Networks for Attention Learning in Fine-grained Action Recognition." in *VISIGRAPP (4: VISAPP)*, 2019, pp. 311–318.

[18] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya, J. Seekins, D. A. Mong, S. S. Halabi, J. K. Sandberg, R. Jones, D. B. Larson, C. P. Langlotz, B. N. Patel, M. P. Lungren, and A. Y. Ng, "CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 590–597, Jul. 2019. [Online]. Available: https://aaai.org/ojs/index.php/AAAI/article/view/3834

[19] J. Shiraishi, S. Katsuragawa, J. Ikezoe, T. Matsumoto, T. Kobayashi, K.-i. Komatsu, M. Matsui, H. Fujita, Y. Kodera, and K. Doi, "Development of a Digital Image Database for Chest Radiographs With and Without a Lung Nodule," *American Journal of Roentgenology*, vol. 174, no. 1, pp. 71–74, Jan. 2000, publisher: American Roentgen Ray Society. [Online]. Available: https://www.ajronline.org/doi/full/10.2214/ajr.174.1.1740071

[20] S. Jaeger, S. Candemir, S. Antani, Y.-X. J. Wáng, P.-X. Lu, and G. Thoma, "Two public chest X-ray datasets for computer-aided screening of pulmonary diseases," *Quantitative Imaging in Medicine and Surgery*, vol. 4, no. 6, pp. 475–477, Dec. 2014. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4256233/

[21] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár, "Microsoft COCO: Common Objects in Context," *arXiv:1405.0312 [cs]*, Feb. 2015, arXiv: 1405.0312. [Online]. Available: http://arxiv.org/abs/1405.0312

[22] L. Oakden-Rayner, "Half a million x-rays! First impressions of the Stanford and MIT chest x-ray datasets," Feb. 2019. [Online]. Available: https://laurenoakdenrayner.com/2019/02/25/half-a-million-x-rays-first-impressions-of-the-stanford-and-mit-chest-x-ray-datasets/