GraphX$^{\mathbf{NET}}$– Chest X-Ray Classification Under Extreme Minimal Supervision

Angelica I. Aviles-Rivero*[1], Nicolas Papadakis ⋆[2], Ruoteng Li[3], Philip Sellars[1], Qingnan Fan[4], Robby T Tan[3,5], and Carola-Bibiane Schönlieb[1]

[1] DPMMS and DAMPT, Faculty of Mathematics, University of Cambridge, UK
{ai323,ps644,cbs31}@cam.ac.uk
[2] IMB, Universite de Bordeaux nicolas.papadakis@math.u-bordeaux.fr
[3] National University of Singapore, Singapore liruoteng@u.nus.edu
[4] Stanford University, USA fqnchina@gmail.com
[5] Yale-NUS College, Singapore robby.tan@yale-nus.edu.sg

Abstract. The task of classifying X-ray data is a problem of both theoretical and clinical interest. Whilst supervised deep learning methods rely upon huge amounts of labelled data, the critical problem of achieving a good classification accuracy when an extremely small amount of labelled data is available has yet to be tackled. In this work, we introduce a novel semi-supervised framework for X-ray classification which is based on a graph-based optimisation model. To the best of our knowledge, this is the first method that exploits graph-based semi-supervised learning for X-ray data classification. Furthermore, we introduce a new multi-class classification functional with carefully selected class priors which allows for a smooth solution that strengthens the synergy between the limited number of labels and the huge amount of unlabelled data. We demonstrate, through a set of numerical and visual experiments, that our method produces highly competitive results on the ChestX-ray14 data set whilst drastically reducing the need for annotated data.

Keywords: Semi-Supervised Learning · Classification · Chest X-Ray · Graphs · Transductive Learning

1 Introduction

The Chest X-Ray (CXR) is the most commonly performed x-ray examination which captures details of the lungs, heart, bones and blood vessels. CXRs play a critical role in diagnosing and monitoring conditions such as pneumonia, heart problems and lung cancer. However, it remains one of the most complex imaging studies to interpret [10]. The effectiveness and accuracy of the interpretation heavily relies on the radiologist's expertise and still there is a substantial clinical error on the outcome [4]. Furthermore, the requirement of human expertise increases the finical cost and time required for evaluation. Therefore, there is a clear need for fast automated evaluations of CXRs.

⋆ Equal Contribution

CXR classification has been widely addressed by the community, yet it remains an open problem. Early developments were based in handcrafted classification e.g. [16]. However, this set of algorithmic approaches require particular modelling hypothesis to be met (e.g. texture, geometry, intensity), which may not be feasible to fulfill in practice. Due to the incredible results produced by deep learning in the field of computer vision, there has been a rush to apply deep learning architectures to the classification of CXRs [19,17,1], which have shown promising results. The majority of these methods utilise deep convolutional neural network with architectures such as ResNet [12], due to the success of these architectures in computer vision classification tasks. Several training methods have been considered including: pre-trained networks, fine tuned networks and networks trained from scratch on X-ray data e.g. [19,17,1].

However, a major drawback of these techniques is the high dependence on a large corpus of labelled data. Particularly in the medical domain, this might be a strong assumption for a solution, as annotated data contains strong human bias. Although there has been a huge effort in the community to mitigate this drawback by providing datasets such as ChestX-ray14, the has annotations but is far from being a definite expression of ground truth [14]. Therefore, by using supervised learning techniques one allows the labelling error and uncertainty to adversely effect the classification output of our machine learn framework. To tackle both the effect of human bias and the limited amount of labelled data, we propose using the power of semi-supervised learning and graph representations.

Our Contributions. We propose a novel semi-supervised graph-based framework called GraphX$^{\mathbf{NET}}$. Our contributions are: 1) *a new multi-class classification functional with carefully chosen class priors.* Our framework is based on the normalised and non-smooth $p = 1$ Laplacian. 2) We demonstrate that our novel framework learns to accurately classify CXRs, with a performance comparable to state-of-the-art deep learning techniques, whilst using an extremely smaller amount of labelled data. 3) This work also represents the first time that graph representations have been used for X-ray classification.

2 GraphX$^{\mathbf{NET}}$ Framework for X-Ray Data Classification.

Our approach is motivated by a central problem in medical imaging which is the lack of reliable quality annotated data. Although, transfer learning [1] or Generative Adversarial Networks [15] somewhat mitigate this problem, they fail to account for the mismatch between expert annotation and ground truth annotation created by human bias and uncertainty. With this motivation in mind, we propose, for the first time, using a semi-supervised framework, call GraphX$^{\mathbf{NET}}$ (see Fig. 1 for illustration).

Data Representation with Graphs. Although there are different methods for representing data including conventional grid form. In this work, we motivate the use of graph data representations as follows. Firstly, graphs are a natural representation for groups of images where each node represents an individual image. Secondly, given that graph based methods seek to find smooth solutions to

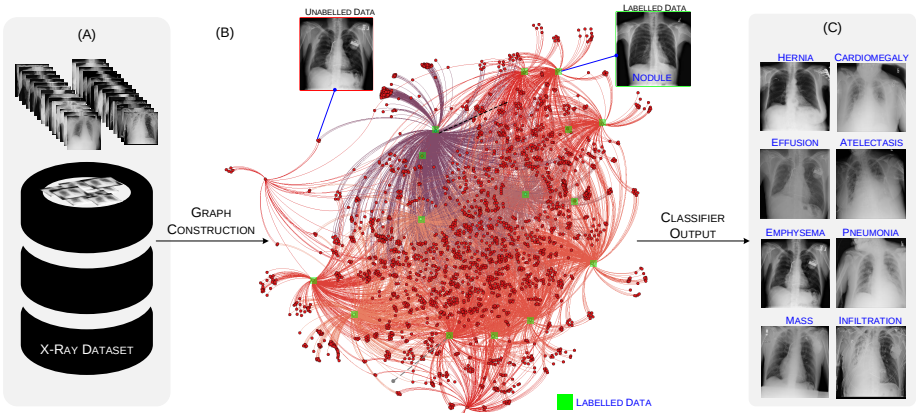


Fig. 1: Overview of our proposed GraphX$^{\mathbf{NET}}$ method. We exploit both labeled and unlabeled data to produce high classification accuracy. In this framework, we aim to propagate labels for the unlabeled data with minimal supervision.

the created embedding, they are able to correct for initially mislabelled samples. Lastly, graph has strong and mathematical properties such as sparseness which allows for fast computation.

We represent a given dataset as an undirected weighted graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, W)$ compromising a set of n nodes $\mathcal{V}$ which are connected by a set of edges $\mathcal{E}$ with weights $w_{ij} = S(i,j) \geq 0$ that correspond to some similarity measure S between the features of nodes $i \in \mathcal{V}$ and $j \in \mathcal{V}$, $w_{ij} = 0$ if $(i,j) \notin \mathcal{E}$; and functions $u \in \mathbb{R}^n$. Our setting is based on the normalised graph p-Laplacian, which reads:

$$\Delta_p(u) = \sum_{i,j} w_{ij} \left\| \frac{u_i}{d_i^{1/p}} - \frac{u_j}{d_j^{1/p}} \right\|^p, \text{ with } p \geq 1 \text{ and } d_i = \sum_j w_{ij} > 0, \quad (1)$$

where d_i is the degree of node i. The eigenfunctions of the graph Laplacian operator give interesting understanding of the substructures of the graph. Eigenfunctions of a normalised graph Laplacian for $p = 2$ have been successfully used in different applications such as in [2,7,8].

Learning to Classify under Extreme Minimal Supervision. However, unlike those works, our framework has a different aim which is to solely obtain classification estimates on the unlabeled samples. That is, to perform a node classification task on $\mathcal{G}$ with L available classes, given an extremely small amount of labelled nodes x_i. More precisely, given a small amount of labeled data $\{(x_i, y_i)\}_{i=1}^{l}$ with provided labels $\mathcal{L} = \{1, .., L\}$ and $\{y_i\}_{i=1}^{l} \in \mathcal{L}$ and a large amount of unlabelled data $\{x_k\}_{k=l+1}^{n}$, we seek to infer a function $f : \mathcal{X}^n \mapsto \mathcal{Y}^n$ such that f gets a good estimate for $\{x_k\}_{k=l+1}^{n}$.

Although several works have explored this learning style, either from a pure machine learning perspective e.g.[20] or a medical imaging perspective e.g. [18],

these methods seek to only approximate $p \to 1$ in the graph Laplacian. However, recent developments on machine learning showed that the use of the unnormalised (i.e. without the re-scaling by the node degrees in (1)) and non smooth $p = 1$ Laplacian, related to total variation, can achieve better performance [5].

To mitigate these current drawbacks in the literature, we propose a novel semi-supervised framework, GraphX$^{\mathbf{NET}}$, based on the normalised and non smooth $p = 1$ Laplacian in (1). The function can then be rewritten as: $\Delta_1(u) = |WD^{-1}u|$, where W is the weight matrix w_{ij} and D the diagonal matrix containing the degrees d_i. To this end, we generalise the unsupervised binary normalised graph method of [9] to a semi-supervised multi-class graph approach. To this aim, our algorithmic approach is as follows.

For each class, $k = 1 \cdots L$, we consider a variable u^k that has values for all nodes of the graph. For all unlabeled nodes $i > l$, the L variables are then coupled with the constraints that for all nodes i: $\sum_{k=1}^{L} u_i^k = 0$, $\forall i > l$. This simple coupling indeed leads to faster projection algorithms than simplex [3,11] or non convex orthogonality constraints between u^k's [8]. We assume that a set of annotated nodes $\mathcal{I}_k \subset \{1 \cdots l\}$ are available for each class k: $y_i = k \in \mathcal{L}$ for all $i \in \mathcal{I}_k$. Taking a small parameter $\epsilon > 0$, we therefore constrain that:

$$\begin{cases} u_i^k \geq \epsilon & \text{if } i \in \mathcal{I}_k \\ u_i^{k'} \leq -\epsilon \text{ if } i \in \mathcal{I}_k \text{ and } k' \neq k. \end{cases} \tag{2}$$

This information is then used in an iterative PDE process with a time parameter t, in which we seek to minimise the sum of normalised ratios $\sum_k \frac{\Delta_1(u^k)}{|u^k|}$. Denoting $\mathbf{u} = [u^1, \cdots u^L]$ and a time step $\Delta t > 0$. Then formally, we seek to minimise:

$$\mathbf{u}^{(t+1)} = \underset{\mathbf{u}}{\operatorname{argmin}} \frac{\|\mathbf{u} - \mathbf{u}^{(t)}\|^2}{2\Delta t} + \sum_{k=1}^{L} \left(\Delta_1(u^k) - \frac{\Delta_1(u^{k,(t)})}{|u^{k,(t)}|} \langle \operatorname{sign}(u^{k,(t)}), u^k \rangle \right), \tag{3}$$

under the set of previously described coupling and data (2) constraints. Following [13,9], a final shifting $u^{k,(t+1)} = u^{k,(t+1)} - \operatorname{median}(u^{k,(t+1)})$ and a normalisation $\mathbf{u}^{(t+1)} = \mathbf{u}^{(t+1)}/||\mathbf{u}^{(t+1)}||$ are necessary at the end of each iteration to prevent from converging to trivial solutions.

When a unique u^k is considered, the scheme iteratively decreases the ratio $\frac{\Delta_1(u^{k,(t)})}{|u^{k,(t)}|}$ since $\langle \operatorname{sign}(u^{k,(t)}), u^{k,(t)} \rangle = |u^{k,(t)}|$, so that the solution $u^{k,(t+1)}$ of (3) necessarily satisfies:

$$\Delta_1(u^{k,(t+1)}) \leq \frac{\Delta_1(u^{k,(t)})}{|u^{k,(t)}|} \langle \operatorname{sign}(u^{k,(t)}), u^{k,(t+1)} \rangle \leq \frac{\Delta_1(u^{k,(t)})}{|u^{k,(t)}|} |u^{k,(t+1)}|. \tag{4}$$

As noticed in [9], the scheme makes $u^{k,(t)}$ converge to a bivalued function that naturally segment the graph. As L variables are coupled, the final labelling of a node i is chosen from the variable u_i^k with the highest value: $y_i = \underset{k}{\operatorname{argmax}}\, u_i^k$.

Optimisation Scheme. For each time step t, the problem (3) is solved at successive time steps using the accelerated primal dual algorithm of [6]. Denoting

as $\mathbf{v} = \mathbf{u}^{(t)}$ the current estimation and initialising $\mathbf{u}_0 = \tilde{\mathbf{u}}_0 = \mathbf{v}$, $z_0^k = WD{-}1u_0^k$, the algorithm to obtain $\mathbf{u}^{(t+1)}$ with an iterative sequence $\mathbf{u}_\ell$ indexed by ℓ reads:

$$\begin{cases} z_{\ell+1}^k = z_\ell^k + \sigma_\ell W D^{-1} \tilde{u}_\ell^k \\ z_{\ell+1}^k = \frac{z_{\ell+1}^k}{\max(1, |z_{\ell+1}^k|)} \\ u_{\ell+1}^k = \frac{u_\ell^k + \tau_\ell \Delta t \left(\frac{\Delta_1(v^k)}{|v^k|} \operatorname{sign}(v^k) + D^{-1} W z_{\ell+1}^k \right)}{1 + \tau_\ell \Delta t} \\ u_{\ell+1}^k = \operatorname{Proj}_C(u_{\ell+1}^k) \\ \gamma_\ell = 1/\sqrt{1 + \tau_\ell / \Delta t},\ \tau_{\ell+1} = \tau_\ell \gamma_\ell,\ \sigma_{\ell+1} = \sigma_\ell / \gamma_\ell \\ \tilde{u}_{\ell+1}^k = u^k + \gamma_\ell (\tilde{u}^{k+1} - u^k), \end{cases}$$

where the projection onto the set of constraints C combining the coupled constraint and (2) reads pointwise:

$$\operatorname{Proj}_C(u_i^k) = \begin{cases} \max(u_i^k, \epsilon) & \text{if } i \in \mathcal{I}_k \\ \min(u_i^k, -\epsilon) & \text{if } i \in \mathcal{I}_{k'} \text{ and } k' \neq k. \\ u_i^k - \frac{1}{L} \sum_{k'} u_i^{k'} & \text{if } i > l. \end{cases} \tag{5}$$

For positive parameters σ_0 and τ_0 satisfying $\sigma\tau < 4$, such process makes $\mathbf{u}_\ell$ converges to $\mathbf{u}^{(t+1)}$, the solution of (3).

3 Experimental Results

This section is devoted to describe in detail the set of experiments that we conducted to validate our GraphX$^{\mathbf{NET}}$ approach.

Data Description. We evaluate our approach using the ChestX-ray14 [17] dataset, which is composed of 112, 120 frontal chest view X-ray with size of 1024× 1024. The dataset is composed of 14 classes (pathologies). All measurements were taken from this dataset.

Evaluation Methodology. We validate our theory as follows. Firstly, we visualise the graphical construction and classification tasks of our graph-based semi-supervised framework. Secondly, the main part of the evaluation is to compare our GraphX$^{\mathbf{NET}}$ to the state-of-the-art methods on X-ray classification. We compare ours against two deep learning techniques: WANG17 [17] and YAO18 [19]. To evaluate the classifier output quality of the compared approaches, we performed a ROC analysis using the area under the curve (AUC) per pathology along with their average. Finally, beside the official split, we perform a comparison with random partitions on ChestX-ray8 using WANG17 [17] as baseline.

Results and Discussion. Firstly, we start by giving some insight into our approach with some visualisations shown in Fig. 2. The left side of the figure shows two graphs in which the first one illustrates the initial state of the graph created after computing the feature distances between the given X-ray data while the second one shows the graph after computing (3). The colours on the graph

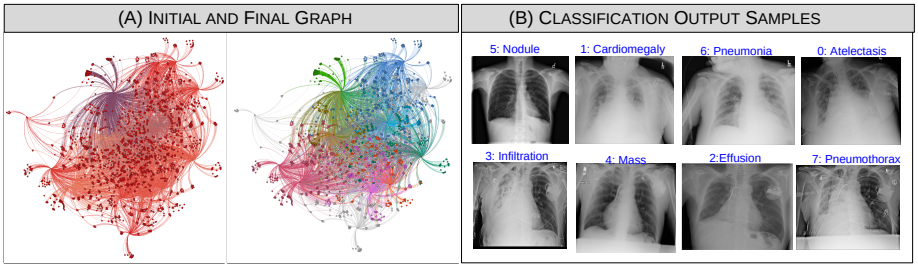
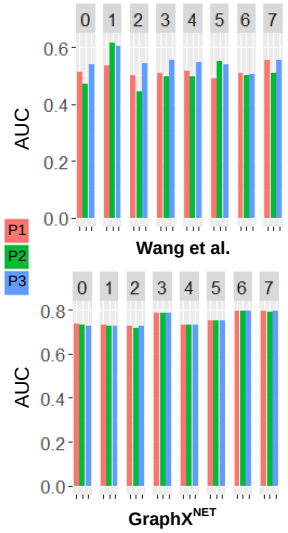


Fig. 2: Graphical Construction and Classification: (A) shows the graphical representation of the ChestX-ray14 dataset, where in the final classified graph, each colour represents a different class and (B) demonstrates examples of correct classifications produced by our framework.

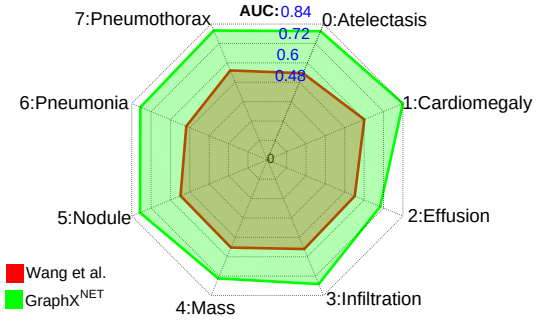


Pathology	Wang17 [17]	Yao18[19]	GraphX$^{\mathbf{NET}}$
Cardiomegaly	0.81	0.856	0.8799
Emphysema	0.833	0.843	0.8407
Edema	0.805	0.806	0.802
Hernia	0.872	0.775	0.8722
Pneumothorax	0.799	0.805	0.837
Effusion	0.759	0.806	0.792
Mass	0.693	0.777	0.809
Fibrosis	0.786	0.743	0.8034
Atelectasis	0.7	0.733	0.7189
Consolidation	0.703	0.711	0.7336
Pleural	0.684	0.724	0.757
Thicken	0.669	0.724	0.7205
Nodule	0.658	0.684	0.7113
Pneumonia	0.661	0.673	0.7664
Average AUC	0.7451	0.7614	0.7888

Table 1: Comparison of the classification accuracy of GraphX$^{\mathbf{NET}}$ against two state-of-the-art deep learning method, Wang et al. [17] and Yao et al. [19]. Here we report the AUC measure over all 14 pathology classes along with the overall average. Plots on the left side highlight the sensitivity of the AUC for each class when changing the data partition of the data set (using 15% for training)

indicates an images belonging to a particular class. The right side shows few sample graph label output, that were correctly classified, of our approach.

To evaluate the performance of our approach, we compared it against state of the art Deep Learning approaches, namely WANG17 [17] and YAO18 [19]. To the best of our knowledge, there are no semi-supervised learning method, for X-ray classification, that we can compare against. Therefore, we set as our



WANG17 [17]					
%Labeled	70%				
AUC	0.548				

GraphX$^{\mathbf{NET}}$					
%Labeled	2%	5%	10%	15%	20%
AUC	0.53	0.58	0.63	0.68	0.78

Table 2: Comparison of the classification accuracy of GraphX$^{\mathbf{NET}}$ against a state-of-the-art deep learning method by Wang et al. [17]. We give the average AUC measure over all eight classes using different amounts of labelled data. Additionally, we give a class by class comparison between the two methods using 70% of the labelled data for the Wang method and 20% for GraphX$^{\mathbf{NET}}$.

baseline WANG17 and YAO18. Table 1 shows the AUC results of the compared approaches where overall our approach outperformed the other methods across most pathology. Even though YAO18 performs better in some classes, a clear advantage of our approach over these two baselines is that while their approach rely in a huge percentage of data, 70%, we were able to report a better average AUC result with only 20% of the data.

Moreover, due to the semi-supervised nature of the GraphX$^{\mathbf{NET}}$ framework, the classification output is very stable with respect to changes in the partition of the dataset. In the plots next to Table 1, we tested the AUC of both the GraphX$^{\mathbf{NET}}$ framework and WANG17 [17] using three different random data partitions, including the partition suggested by Wang. The Wang method is very sensitive to changes in the partition due to the face that supervised methods are heavily reliant on the training set being representative. However, there is minimal change in the performance of GraphX$^{\mathbf{NET}}$ over the three different partitions as the underlying graphical representation is invariant to the partition.

For more detailed analysis of this dependency on the portioning and to further support the advantage of our GraphX$^{\mathbf{NET}}$, in Table 2, we compare the AUC produced by GraphX$^{\mathbf{NET}}$ against WANG17 using a random split over ChestX-ray8. We find that GraphX$^{\mathbf{NET}}$ produces a more accurate classification using 5% of the data labels than the WANG17 method does using 70% of the data labels. Furthermore, as we feed GraphX$^{\mathbf{NET}}$ more of the data labels, the classification accuracy increases and becomes competitive against other the deep learning framework of that YAO18 [19] whilst using a far smaller amount of data labels.

4 Conclusion

In this work, we tackled the problem of X-ray classification and introduced a novel semi-supervised framework based on a graph-based optimisation model, which is the first method that exploits graph-based semi-supervised learning for X-ray data classification. We also introduced a new multi-class classification functional with carefully selected class priors that allows for a smooth solution. We demonstrated that our method produces highly competitive results on the ChestX-ray14 data set whilst drastically reducing the need for annotated data.

Acknowledgments

AIAI is supported by the CMIH, University of Cambridge. NP is supported by the European Unions Horizon 2020 research and innovation programme under the Marie Skodowska-Curie grant No 777826. CBS acknowledges Leverhulme Trust (Breaking the non-convexity barrier), the Philip Leverhulme Prize, the EPSRC grants EP/M00483X/1 and EP/N014588/1, the European Union Horizon 2020, the Marie Skodowska-Curie grant 777826 NoMADS and 691070 CHiPS, the CCIMI and the Alan Turing Institute.

References

1. Bar, Y., Diamant, I., Wolf, L., Lieberman, S., Konen, E., Greenspan, H.: Chest pathology detection using deep learning with non-medical training. In: International Symposium on Biomedical Imaging (ISBI). pp. 294–297 (2015)
2. Belkin, M., Niyogi, P.: Laplacian eigenmaps for dimensionality reduction and data representation. Neural Computation pp. 1373–1396 (2003)
3. Bresson, X., Laurent, T., Uminsky, D., Von Brecht, J.: Multiclass total variation clustering. In: Advances in Neural Information Processing Systems (2013)
4. Bruno, M.A., Walker, E.A., Abujudeh, H.H.: Understanding and confronting our mistakes: the epidemiology of error in radiology and strategies for error reduction. Radiographics **35**(6), 1668–1676 (2015)
5. Bühler, T., Hein, M.: Spectral clustering based on the graph p-laplacian. International Conference on Machine Learning (ICML) (2009)
6. Chambolle, A., Pock, T.: A first-order primal-dual algorithm for convex problems with applications to imaging. Journal of mathematical imaging and vision (2011)
7. Chen, H., Li, K., Zhu, D.e.a.: Inferring group-wise consistent multimodal brain networks via multi-view spectral clustering. IEEE Transactions on Medical Imaging (TMI) pp. 1576–1586 (2013)
8. Dodero, L., Gozzi, A., Liska, A., Murino, V., Sona, D.: Group-wise functional community detection through joint laplacian diagonalization. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (2014)
9. Feld, T.M., Aujol, J.F., Gilboa, G., Papadakis, N.: Rayleigh quotient minimization for absolutely one-homogeneous functionals. Inverse Problems (2019)
10. Folio, L.R.: Chest imaging: an algorithmic approach to learning. Springer (2012)

11. Gao, Y., Adeli-M, E., Kim, M., Giannakopoulos, P., Haller, S., Shen, D.: Medical image retrieval using multi-graph learning for mci diagnostic assistance. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). pp. 86–93 (2015)
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
13. Hein, M., Setzer, S., Jost, L., Rangapuram, S.S.: The total variation on hypergraphs-learning on hypergraphs revisited. In: Advances in Neural Information Processing Systems (2013)
14. Kohli, M.D., Summers, R.M., Geis, J.R.: Medical image data and datasets in the era of machine learningwhitepaper from the 2016 c-mimi meeting dataset session. Journal of digital imaging pp. 392–399 (2017)
15. Moradi, E., Pepe, A., Initiative, A.D.N., et al.: Machine learning framework for early mri-based alzheimer's conversion prediction in mci subjects. Neuroimage (2015)
16. Toriwaki, J.I., Suenaga, Y., Negoro, T., Fukumura, T.: Pattern recognition of chest x-ray images. Computer Graphics and Image Processing pp. 252–271 (1973)
17. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M.: Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2097–2106 (2017)
18. Wang, Z., Zhu, X., et al.: Progressive graph-based transductive learning for multimodal classification of brain disorder disease. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (2016)
19. Yao, L., Prosky, J., Poblenz, E., Covington, B., Lyman, K.: Weakly supervised medical diagnosis and localization from multiple resolutions. arXiv preprint arXiv:1803.07703 (2018)
20. Zhu, X., Ghahramani, Z., Lafferty, J.D.: Semi-supervised learning using gaussian fields and harmonic functions. In: International conference on Machine learning (ICML). pp. 912–919 (2003)