Deep Learning for the Automatic Detection of Congenital Lung Abnormalities

Shaimaa Bakr
Stanford University
sbakr@stanford.edu

David Xue
Stanford University
dxue@cs.stanford.edu

Dominic Abbondanzo
Stanford University
dabbonda@stanford.edu

Mazin Bokhari
Stanford University
mazin@stanford.edu

1. Introduction

Diagnosis of congenital lung abnormalities antenatally allows physicians to (a) be aware of potential management issues during or after delivery, and (b) provide parents with information on the prognosis. The common congenital lung abnormalities can be categorized into three broad classes: bronchopulmonary anomalies, vascular anomalies, and combined lung and vascular anomalies. Fetal MRI is an invaluable diagnostic tool complementary to ultrasound thanks to its high contrast and resolution. Clinically, ultrasound is the first standard-of-care tool to monitor fetal development. Abnormal findings on ultrasound are indications for further investigation through MRI to obtain a precise diagnosis and quantification of compromised development. Fetal MRI volumetry, signal intensities and tissue contrast contain important information on lung growth, maturation and structure of the fetal lung. Although congenital lung abnormalities are rare diseases, for example, congenital pulmonary airway malformation (CPAM) occurs in 1 out of 30,000 pregnancies, precise diagnosis and prognosis are crucial to supporting physicians in clinical management and to informing patients with likely outcomes. The development of fast and fully automatic classification models that require no expert knowledge to pre-process improves clinical decision-making and provides an entry point to regression models of lung volumetry to further improve prognostic prediction.

2. Related Work

Over the last decade, the ability of computer programs to extract information from images has increased tremendously. We owe most of this advancement to convolutional neural networks (CNNs), a type of neural network specialized for processing image data. CNNs have consistently outperformed classical machine learning (ML) techniques since 2012, when AlexNet won the ImageNet Large Scale Visual Recognition Competition [7], a deep neural network takes raw input (possibly after some preprocessing) and automatically learns features through training. In the last few years, we have seen how even better results can be obtained with deep learning [16].

Convolutional neural networks has proven to be very successful in not just natural image classification, but also medical image classification and segmentation. Machine learning has became the dominant technology for tackling computer-aided diagnosis (CAD) in the lungs, generally producing better results than classical rule-based approaches. CAD has been used for other types of diagnostic tasks: breast cancer localization by GoogLeNet and skin cancer classification [1, 17].

For many problems in medical imaging, the information which is necessary for discriminating abnormal from normal cases can be minuscule in comparison to the total complete image and sometimes is present only in a small subset of the data. One way to tackle this is by using Gaussian Mixture Models [8] to pinpoint viable regions, however this is still an open and important problem.

3. Data

The data set consists of 1591 fetal Single Shot Fast-Spin Echo (SSFSE) T2-weighted MRI volumetric scans. The scans were collected during the period between 2004-2017 from the Stanford School of Medicine. Most commonly, patients received routine ultrasound resulting in abnormal findings that indicated further MRI investigations to obtain a precise diagnosis or quantitative information such as lung volume. In the case of high risk of abnormality fetuses, MRI indicated even in the case of normal ultrasound findings. Each scan is labeled as abnormal or normal (1212 normal and 379 abnormal). Labels were collected from the associated radiology report and reviewed by a pediatric radiologist with twelve years experience to verify the diagnosis. The image slices in each are 256 x 256 pixels with an

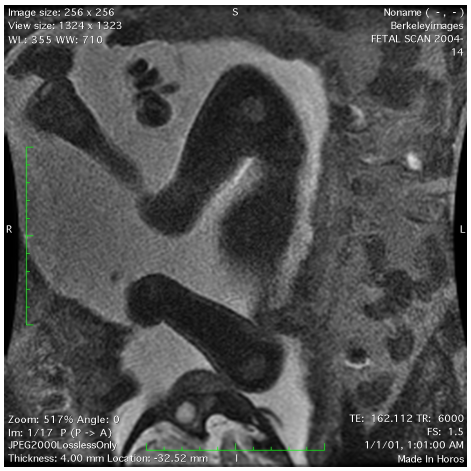

Figure 1. A T2-weighted, single-shot, fast spin echo sagittal MRI slice of the fetus.

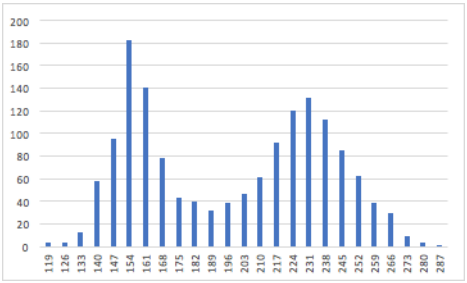

Figure 2. Histogram of Gestational Age. Gestational Age in days at MRI scan date calculated from fetal unltraonography data

average 4.4mm pixel size. The pixel values are grayscale.

There is wide variability in gestational age and stages of fetal lung development. Some patients may have more than one scan associated for follow-up reasons; in this case each scan will be treated as a separate patient for purposes of training and prediction.

3.1. Types of lung abnormalities

As described in the previous section there are many types of lung abnormalities. In this subsection, we provide a brief overview of these types, with more focus on abnormalities available in our data set.

3.1.1 Bronchopulmonary abnormalities

In our data set the the most common anomalies are of this type. Bronchopulmonary abnormalities are specific to the lung bud and further include several conditions that lead to pulmonary underdevelopment. These are characterized by absence of lung tissue or presence of small airways and other lung structures. Other bronchopulmonary abnormalities are congenital pulmonary airway malformation (CPAM), congenital lobar overinflation and bronchogenic cysts.

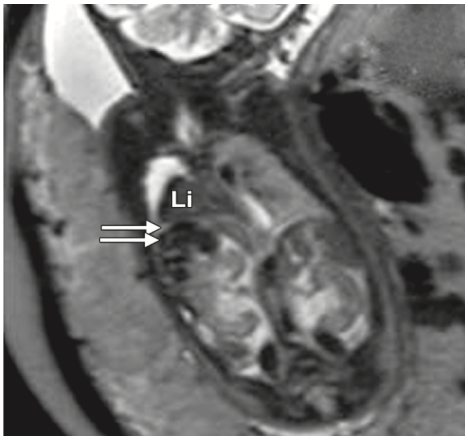

Figure 3. A T2-weighted showing left sided CDH with liver herniated into the thorax.

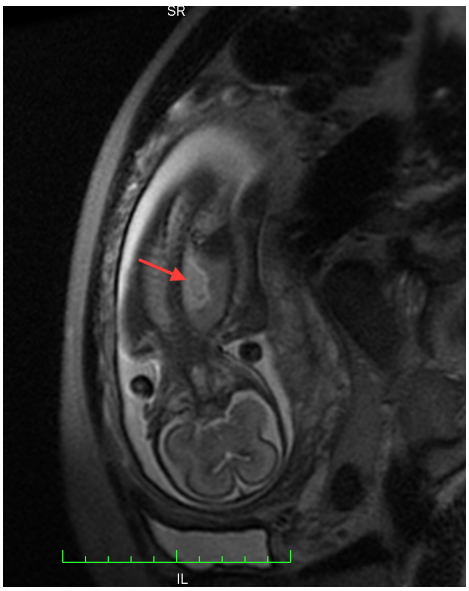

Figure 4. A T2-weighted, single-shot, fast spin echo showing CPAM example

Specifically, pulmonary underdevelopment caused by congenital diaphragmatic hernia and congenital pulmonary airway malformation are the most encountered abnormalities on our data set.

In Congenital Diaphragmatic Hernia (CDH), we observe abdominal structures such as liver or stomach in an intrathoracic position causing compression of thoracic structures such as the lungs and heart.

In congenital pulmonary airway malformation, we observe lung lesions that develop as a result of airway maldevelopment. CPAMs are classified into five types (0-4) based on their airway origin: tracheal, bronchial, bronchiolar, alveolar, or distal acinar.

3.1.2 Vascular abnormalities

Vascular abnormalities include absence of the main pulmonary artery, anomalous origin of the left pulmonary artery and anomalous pulmonary venous drainage.

3.1.3 Combined lung and vascular anomalies

Combined lung and vascular anomalies include scimitar syndrome and bronchopulmonary sequestration.

4. Methods

4.1. Preprocessing

4.1.1 Normalizations

Due to the nature of MRI, even images of the same patient on the same scanner at different can have different intensities. Many MRI models use an intensity normalization from Nyul et al. [11] to alleviate this problem. Additionally, as is typical with CNNs, each input channel (i.e. sequence) is normalized to have zero mean and unit variance within the training set. All images will be preprocessed with histogram equalization to increase contrast within each MRI image. Each image will have some random small amount of Gaussian noise added to each pixel value.

4.1.2 Data Augmentation

Medical imaging data are not readily available in large quantities. Many difficulties lead to scarcity of medical imaging data sets. First, regulations related to patient privacy require extra processing steps to anonymize the patient data before the data can be used for research purposes. Secondly, the incidence of a medical condition (probability of occurrence of a given medical condition in a population within a specified period of time) limits the number of scans produced related to this disease. Lastly, labeling medical image data for research purposes requires the effort and time of expert radiologist, pathologists or other medical experts and thus represents a significant cost and can be a bottleneck in the data set building pipeline. In contrast, natural images are easily produced and shared publicly on a daily basis, and are available for curation and labeling by researchers and through crowd-sourcing efforts. Even for the most common diseases, most publicly-available medical imaging data sets contain hundreds of images, whereas ImageNet has 14 million. Data augmentation is mainly employed to increase the training samples to mitigate overfitting. It is a common practice to use data augmentation in computer vision tasks in which (i) the CNN architectures are very deep, and (ii) obtaining enormous amounts of labeled training data is difficult. Furthermore, the orientation of the fetus is arbitrary. Each training image, before being input into a neural network, will be rotated 0, 90, 180, or 270 degrees. Additionally, each image will be reflected at the mid-sagittal plane.

4.2. Network Architectures

4.2.1 2D-CNN

Several approaches to transforming 3D medical images into 2D images exist in the literature and aim to reduce to dimensionality and thus complexity of the network as well as benefit from existing successful 2D architecture and wide availability of 2D natural image data sets.

Transfer learning is the use of pre-trained networks to try to work around the requirement of large data sets for deep network training; models for processing medical images have greatly benefited from pre-training on natural image data sets like ImageNet [9]. They have also been shown to perform better if pre-trained on other medical imaging data which adapts them to better leverage the intrinsic structure of medical imaging.

We will explore successful 2D pre-trained models such as DenseNet [5] or ResNet [4], using our data set to fine-tune these networks. It is a non-trivial task to identify how to best use expert knowledge as a prior for the model, and if its effect is going to be positive or negative on the model.

Our next approach is to view the 3D data as a sequence of 2D images and employ a sequential model to the 2D input slices. For this future direction, a recent work from Monika Grewal et al. from Parallel Dots [2], published on Jan 2018, described such model which they called RADNet (Recurrent Attention DenseNet). It uses a DenseNet architecture to extract features, in addition to passing sequential data through a bi-directional LSTM layer. It uses the context around each image in the series to make better predictions.

4.2.2 3D-CNN

The same principles and architectures can be extended to three dimensions to obtain 3D-CNNs that are suitable for volumetric data. Authors have used different approaches to integrate 3D in an effective manner with custom architectures [3, 14, 6]. But because of the extra dimension, 3D convolutional networks are more memory intensive than 2D networks. In a 3D convolutional network, it is not only the input image that is larger, but also the representations after each layer in the network. These image representations need to be cached for back propagation, consuming extensive memory. Moreover, the added dimension in 3D convolutional networks adds exponentially to the number of parameters needed to train the network.

VoxNet [10] work goes in depth on using 3D convolutional networks for deep representations of 3D volumetric

point cloud input data for object recognition and classification tasks. This provides a good baseline for working on 3D data.

OctNet [13] hierarchically and dynamically partitions the input into sections of different sizes, based on the amount of detail that they contain. This can potentially pay more attention to intricate sections of the volume containing more details, which could be helpful in some tasks like detecting nodules and lesions.

5. Experiment

5.1. Evaluation Metrics

Our first approach in terms of problem definition is to group all congenital lung abnormalities into one class and train our network to classify normal lung versus abnormal lung in fetal MRI. Our data set is imbalanced; the breakdown of the entire data set is around 75:25 normal to abnormal. Two techniques to deal with class imbalance are oversampling and weighted cross entropy loss. Because normal (negative samples) are much more common, the abnormal images (positive samples) are oversampled during training.

The following performance measures can give more insight into the accuracy of our model than traditional classification accuracy: confusion matrix, precision, recall, F1 score (a weighted average of precision and recall), and AUROC score. We can also still use plots of accuracy and loss curves to understand how certain variables (e.g. complexity of the network, input data size) changes the performance of different models over the same training/validation data.

5.2. Baseline

We performed a 70-15-15 train-dev-test set split of the 3D data and then split each 3D matrix into 2D grayscale slices. We pass each grayscale slice image through 3 layers of conv-bn-max_pool-relu, followed by flattening the image and then applying 2 fully connected layers. The output is a `log_softmax` over the 2 labels for each example in the batch. We use `log_softmax` since it is numerically more stable than first taking the softmax and then the log. We use negative loss likelihood since the output is already softmaxed and log-ed.

6. Results

6.1. Baseline

Table 1. Baseline Hyperparameter Search

Learning Rate	**Accuracy**	**Loss**
0.01	0.627622	0.692386
0.001	0.628934	1.09562
0.0001	0.625	0.783576

We performed a hyperparameter search with learning_rate = 0.01, 0.001, and 0.0001. We found that we are able to overfit to the training set over 10 epochs with accuracy of 0.969 and a learning_rate of 0.01. However, our network does not generalize and with an evaluation accuracy of 0.689.

7. Contributions

Dominic: Made script to decompress all DICOM files and organize them in one place to be fed into model.

Mazin and Shaimaa: Processed the original DICOM files to form the data set, including decompression and formation of Numpy structs, as well as normalization of the data set.

David: Wrote scripts to (1) process all decompressed DICOM files into 3D numpy matrices and save to disk, (2) split 3D numpy matrix files into training, dev, and validation sets, (3) convert to 2D normalized numpy slices and save to disk, and (4) read the 2D normalized slices into the PIL data types. Modified PyTorch DataSet class to handle fetal dataset. Ran the full experiment pipeline.

Dominic and David: Modified the baseline PyTorch model to handle 2D grayscaled files and binary classification.

Shaimaa and David: Expanded the proposal to describe data and methods: current progress, results, future work.

7.1. Code Base

Our public Github repository can be found here.

8. Conclusion

9. Acknowledgements

David Xue is sharing this project with the course CS231A.

References

[1] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115, 2017.

[2] M. Grewal, M. M. Srivastava, P. Kumar, and S. Varadarajan. Radnet: Radiologist level accuracy using deep learning for hemorrhage detection in ct scans. *arXiv preprint arXiv:1710.04934*, 2017.

[3] S. Hamidian, B. Sahiner, N. Petrick, and A. Pezeshk. 3d convolutional neural network for automatic detection of lung nodules in chest ct. In *Medical Imaging 2017: Computer-Aided Diagnosis*, volume 10134, page 1013409. International Society for Optics and Photonics, 2017.

[4] 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*, pages 770–778, 2016.

[5] G. Huang, Z. Liu, K. Q. Weinberger, and L. van der Maaten. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, volume 1, page 3, 2017.

[6] X. Huang, J. Shan, and V. Vaidya. Lung nodule detection in ct using 3d convolutional neural networks. In *Biomedical Imaging (ISBI 2017), 2017 IEEE 14th International Symposium on*, pages 379–383. IEEE, 2017.

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

[8] R. Li, R. Perneczky, I. Yakushev, S. Foerster, A. Kurz, A. Drzezga, S. Kramer, A. D. N. Initiative, et al. Gaussian mixture models and model selection for [18f] fluorodeoxyglucose positron emission tomography classification in alzheimers disease. *PloS one*, 10(4):e0122731, 2015.

[9] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak, B. van Ginneken, and C. I. Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017.

[10] D. Maturana and S. Scherer. Voxnet: A 3d convolutional neural network for real-time object recognition. In *Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on*, pages 922–928. IEEE, 2015.

[11] L. G. Nyúl, J. K. Udupa, and X. Zhang. New variants of a method of mri scale standardization. *IEEE transactions on medical imaging*, 19(2):143–150, 2000.

[12] S. Pereira, A. Pinto, V. Alves, and C. A. Silva. Brain tumor segmentation using convolutional neural networks in mri images. *IEEE transactions on medical imaging*, 35(5):1240–1251, 2016.

[13] G. Riegler, A. O. Ulusoy, and A. Geiger. Octnet: Learning deep 3d representations at high resolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 3, 2017.

[14] A. A. A. Setio, F. Ciompi, G. Litjens, P. Gerke, C. Jacobs, S. J. van Riel, M. M. W. Wille, M. Naqibullah, C. I. Sánchez, and B. van Ginneken. Pulmonary nodule detection in ct images: false positive reduction using multi-view convolutional networks. *IEEE transactions on medical imaging*, 35(5):1160–1169, 2016.

[15] N. J. Tustison, B. B. Avants, P. A. Cook, Y. Zheng, A. Egan, P. A. Yushkevich, and J. C. Gee. N4itk: improved n3 bias correction. *IEEE transactions on medical imaging*, 29(6):1310–1320, 2010.

[16] B. van Ginneken. Fifty years of computer analysis in chest imaging: rule-based, machine learning, deep learning. *Radiological physics and technology*, 10(1):23–32, 2017.

[17] D. Wang, A. Khosla, R. Gargeya, H. Irshad, and A. H. Beck. Deep learning for identifying metastatic breast cancer. *arXiv preprint arXiv:1606.05718*, 2016.

A. Appendix

A.1. MRI Imaging

At a high level, MRI works by measuring the radio waves emitting by atoms subjected to a magnetic field. The appearance of tissue in an MRI depends on the tissues chemical composition and which particular MR sequence is employed. In a T2-weighted MRI tissues with more water or fat appear brighter due to their relatively high number of hydrogen atoms. In contrast, bone (as well as air) has low signal and appears dark on T2-weighted images.

A.2. Data File Format

A DCM file is an image file saved in the Digital Imaging and Communications in Medicine (DICOM) image format and is part of the DICOM standard for storing and transmitting medical image data. It stores a medical image, such as a CT scan or ultrasound, and may also contain information about the patient. We use the Grassrooms DICOM (GDCM) library to decompress each DCM file and the PyDicom and NumPy libraries to reconstruct 3D matrix information for each scan.

A.3. Imaging Planes

Each MRI scan acquires images along one of three planes: axial, coronal, and sagittal.

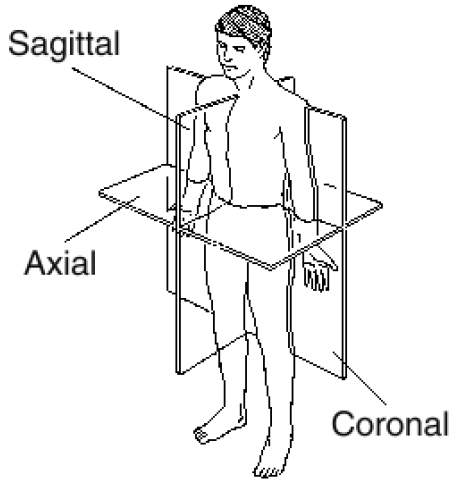

Figure 5. Axial, sagittal, coronal views of a human body

A.4. Registration

If the patient moves during an MRI screening, images may be offset from one another. Intrinsic fetal motion can degrade image quality and thereby introduce motion artifacts and other unwanted effects such as a reduced volumetric precision. If different sequences are combined in a single channel, or if a 3D network is used, then the images must first be aligned to a common orientation.

A.5. Bias field correction

MRI images are affected by bias field distortion, which causes the intensity to vary even across the same tissue [12]. The N4ITK method [15] is the most common method for correcting this.