Deep Learning for the Classification of Congenital Lung Abnormalities using Volumetric MRI Scans

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

Abstract

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. Fetal MRI is an invaluable diagnostic tool providing volumetry, signal intensities and tissue contrast contain important information on lung growth, maturation and structure of the fetal lung. Currently there are no fast and fully automatic classification models of normal versus abnormal fetal lung MRI scans. We explore the application of Convolutional Networks with 2D and 3D kernels to solve this problem.

1. Introduction

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 common congenital lung abnormalities can be categorized into three broad classes: bronchopulmonary anomalies, vascular anomalies, and combined lung and vascular anomalies [11]. Fetal MRI is complementary to ultrasound thanks to its high contrast and resolution [2]. Clinically, ultrasound is the first standard-of-care tool to monitor fetal development [23]. Abnormal findings on ultrasound are indications for further investigation through MRI to obtain a precise diagnosis and quantification of compromised development. 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 [3] to further improve prognostic prediction. We apply different Convolutional Neural Network architectures to medical MRI image volumes to obtain a binary class prediction of normal or abnormal fetal lung.

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 [13], 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 [26].

CNNs has proven to be very successful in not just natural image classification, but also medical image classification and segmentation. Machine learning has become the dominant technology for tackling computer-aided diagnosis (CAD) in the lungs, generally producing better results than classical rule-based approaches [20]. CAD has been used for other types of diagnostic tasks: breast cancer localization by GoogLeNet and skin cancer classification [4, 27]. Most work on applying CNNs to medical imaging has focused on developing deep learning architectures for segmentation tasks, such as U-Net [22] for 2D scans or V-Net [17] for volumetric data. Less work has been done in classification of volumetric medical data, likely due to its challenges sparsity and computational expense.

Medical imaging data is not readily available in large quantities. 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 [15]. 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. Models such as DenseNet [9] and ResNet [8] with simple and effective structures to allow deeper networks have been shown to be successful on a variety of 2D image tasks using principles of transfer learning to particular task domains.

In theory 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 [6, 24, 10]. 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.

Due to these limitation, most work on using 3D-CNNs has explored relatively shallow networks. Hara et al. [7] explored the use of a deep 3D ResNet architecture in the domain of action recognition. The use of very deep 2D CNNs trained on Imagenet has generated outstanding progress in image recognition; the goal of 3D ResNet was to see if the Kinetics dataset [12], a dataset of 300,000 trimmed videos covering 400 categories, could generate similar progress in computer vision for videos.

3. Dataset

The data set consists of 4632 Single Shot Fast-Spin Echo (SSFSE) T2-weighted fetal 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 normal (3485) or abnormal (1147). Labels were collected from the associated radiology report and reviewed by a pediatric radiologist with twelve years experience to verify the diagnosis. We performed a 70-15-15 train-dev-test set split of the 3D data.

There is a wide variability in gestational age and stages of fetal lung development 2. Additionally 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.

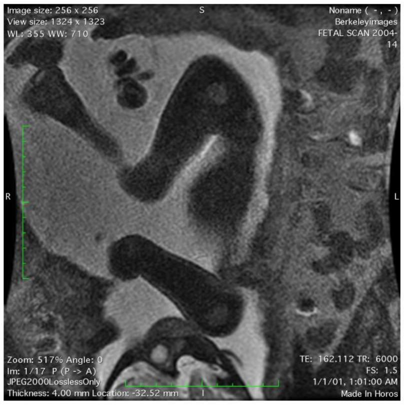

Figure 1. A T2-weighted, single-shot, fast spin echo sagittal MRI slice of the fetus.

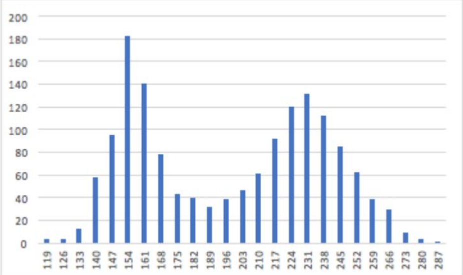

Figure 2. Histogram of Gestational Age. Gestational Age in days at MRI scan date calculated from fetal unltraonography data

3.1. Types of lung abnormalities

Below is a breakdown of abnormalities by structural type. 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. We provide a brief overview of these types, with more focus on abnormalities available in our data set in the appendix.

1. Bronchopulmonary abnormalities
    - Congenital Diaphragmatic Hernia
    - Congenital pulmonary airway malformation
2. Vascular abnormalities
    - Absence of the main pulmonary artery
    - Anomalous origin of the left pulmonary artery
    - Anomalous pulmonary venous drainage
3. Combined lung and vascular anomalies
    - Scimitar syndrome
    - Bronchopulmonary sequestration

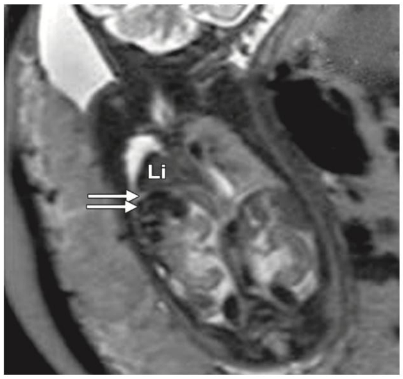

Figure 3. A T2-weighted showing left sided CDH with liver herniated into the thorax.

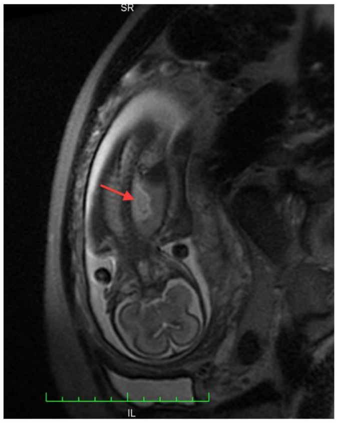

Figure 4. A T2-weighted, single-shot, fast spin echo showing CPAM example

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. The pixel values are a single intensity channel which we converted to grayscale. Many MRI models use an intensity normalization from Nyul et al. [18] 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 were 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

Many difficulties lead to scarcity of medical imaging data sets. First, patient privacy regulations require extra processing steps to anonymize the patient data. Secondly, the incidence of a medical condition (probability of occurrence of a given medical condition) limits the number of scans produced related to this disease. Lastly, labeling medical image data for research purposes requires the effort of expert radiologists,and thus represents 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 crowdsourcing 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 enough labeled training data is difficult. Furthermore, the orientation and size of the fetus can change substantially and we want all orientations to be accounted for by the models. Each scan randomly had one or more of the following augmentations applied before being saved as a new scan: rotations, cropping, random noise, and shearing each with about a probability of 0.3.

4.1.3 Variable Scan Dimensions

Each scan is composed of a variable number of image slices as depicted in Figure 5. Image slices are square with an average 4.4mm pixel size. 4184 of the image slices are of dimension 512 pixels and 429 are dimension 256 pixels. To get the same size to get fixed inputs for our models we truncated scans and padded scans with masking and also downsampled all image slices.

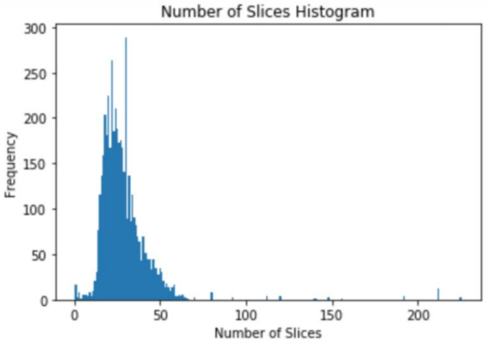

Figure 5. Histogram of Number of Slices per Scan.

4.1.4 Class Imbalance

Our data set is imbalanced; the breakdown of the entire data set is around 70:30 normal to abnormal. Two techniques to deal with class imbalance are oversampling and weighted cross entropy loss.

4.2. Network Architectures

4.2.1 2D-CNN

We implemented a 2D-CNN multi-slice architecture based on the 2D models DenseNet-121 [9] and ResNet-18 [8]. DenseNet-121, shown in Table 1 and ResNet-18 shown to be successful on 2D image tasks. Pre-trained models on ImageNet are available for both architectures. A 3D scan volume is input to the network, as a sequence of slices. Each slice is applied to the 2D DenseNet or ResNet network. Output in the last fully-connected layer are averaged and used to compute an overall prediction per scan as showing in Figure 6.

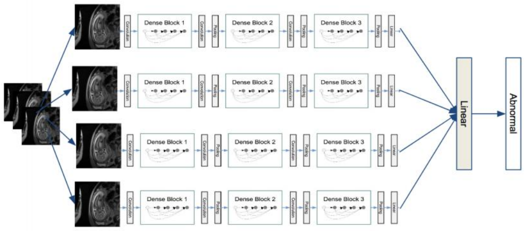

Figure 6. Multi-slice 2D CNNs

We performed two sets of experiments. Firstly, we trained the network from scratch. In the second set of experiments our data set was used to fine-tune these networks. Models with 2D convolutional kernels have the advantage of fewer parameters, which affects training time and reduces overfitting.

Table 1. DenseNet (121-Layer) Architecture.

Layer	Architecture
Convolution	7×7 conv stride 2
Pooling	7×7 max pool stride 2
Dense Block (1)	$\begin{bmatrix} 1 \times 1 conv \\ 3 \times 3 conv \end{bmatrix} \times 6$
Transition Layer (1)	1×1 conv 1×1 avg pool stride 2
Dense Block (2)	$\begin{bmatrix} 1 \times 1 conv \\ 3 \times 3 conv \end{bmatrix} \times 12$
Transition Layer (2)	1×1 conv 1×1 avg pool stride 2
Dense Block (3)	$\begin{bmatrix} 1 \times 1 conv \\ 3 \times 3 conv \end{bmatrix} \times 24$
Transition Layer (3)	1×1 conv 1×1 avg pool stride 2
Dense Block (4)	$\begin{bmatrix} 1 \times 1 conv \\ 3 \times 3 conv \end{bmatrix} \times 16$
Classification Layer	7×7 global average pool 1000D fully-connected softmax

4.2.2 3D-CNN

We followed the architecture of 3D-ResNet18 in Hara et al [7] shown in Table 2 to extend an open source implementation [1] and tested it on our own dataset. Residual blocks are shown in brackets. The difference between our networks and original ResNets [7] is the number of dimensions of convolutional kernels and pooling. 3D ResNets perform 3D convolution and 3D pooling where the sizes of convolutional kernels are $3 \times 3 \times 3$.

Table 2. 3DResNet (18-Layer) Architecture.

Layer Name	Architecture
conv1	$7 \times 7 \times 7, 64$, stride 1
conv2_x	$3 \times 3 \times 3$ max pool, stride 2 $\begin{bmatrix} 3 \times 3 \times 3, 64 \\ 3 \times 3 \times 3, 64 \end{bmatrix} \times 2$
conv3_x	$\begin{bmatrix} 3 \times 3 \times 3, 128 \\ 3 \times 3 \times 3, 128 \end{bmatrix} \times 2$
conv4_x	$\begin{bmatrix} 3 \times 3 \times 3, 256 \\ 3 \times 3 \times 3, 256 \end{bmatrix} \times 2$
conv5_x	$\begin{bmatrix} 3 \times 3 \times 3, 512 \\ 3 \times 3 \times 3, 512 \end{bmatrix} \times 2$

Each convolutional layer is followed by batch normalization and ReLU. Downsampling is performed by conv3_1, conv4_1, conv5_1 with a stride of 2. As a final step we apply an average pool, use a 2-dimension last fully-connected layer set for our dataset (2 categories), and apply a softmax.

5. Experiment

5.1. Evaluation Metrics

Because the data set has a class imbalance, we focus on the following metrics: confusion matrix, precision, recall, F1 score (a weighted average of precision and recall).

5.2. Results

5.2.1 Baseline

As a simple baseline sanity check to process our data, we split each 3D matrix into 2D grayscale slices and pass each 2D slice 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 softmax-ed and log-ed.

We performed a manual hyperparameter search with different learning rates and 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 0.659. These results are to be expected since the slices are each treated independently when in reality each slice may contain no pertinent information to the label.

Table 3. Training Performance

Model	Precision	Recall	F1
ResNet-18	0.978	0.978	0.978
DenseNet-121	0.976	0.889	0.930
3D ResNet-18	1.00	0.992	0.996

5.2.2 2D and 3D-CNN

Precision, recall and F1 for training are reported in Table 3. The 2D experiments were performed on DenseNet-121 and ResNet-18 multi-slice architectures. Training was done both from scratch and on pre-trained model. Images were resized 64x64x40 and 224x224x40 respectively with a batch size of 16 to fit the model onto our 12GB NVIDIA Titan XP GPU. Corresponding image normalization was applied to images before the pre-trained models. We experimented with learning rates 0.001, 0.0001 and 0.00001 using Adam optimizer, with dropout equal to 0.2, 0.3 and 0.4. The confusion matrices for training is showing in Table 4

Table 4. Densenet-121 Training Confusion Matrix

	Predicted Negative	Predicted Positive
True Negative	163	1
True Positive	5	40

For 3D-ResNet model, we downsampled and padded each scan to a volume of 64x64x40 with a batch size of 16 to fit the model onto our 12GB NVIDIA Titan XP GPU. Training was slow due to the low batch size required to fit each scan into GPU and each epoch took approximately 10 minutes. After testing several learning rates, we used Adam optimizer with an initial learning rate of 0.001. To help with overfitting we tried to apply weight decay with a rate of 0.999 and dropout to all layers with a rate of 0.2. The confusion matrices for validation is showing in Table 5

Table 5. 3D ResNet-18 Validation Confusion Matrix

	Predicted Negative	Predicted Positive
True Negative	56	3
True Positive	17	4

5.3. Discussion

We found that our models tended to overfit the training set after around 200 epochs. We found that our regularization and data augmentation had little effect on improving the dev and test set performance. We examined some of the misclassified scans slice by slice and found that many had motion blurred images because of fetal motion.

A significant challenge of classification of volumetric MRI data is that 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 image slices. For our 3D-ResNet the lack of pretrained models like those using ImageNet is a major impediment; however, Hara et al. recently released pretrained models on their Kinetics action recognition video dataset that we are planning to test.

As a further refinement, we take advantage of the localization of abnormality region inside the body of the fetus. Having a two stage model where first the fetus is recognized and bounded and then fed into the model should help performance by discarding irrelevant data. One way to tackle this is by using advanced featurization (e.g. SIFT, HoG) and Gaussian Mixture Models [14] to pinpoint viable regions, however this is still an open and important problem. This could even be taken one step further, where the lung within the fetus is also identified before being fed into the model.

6. Future Work

We plan on obtaining segmentation labels as well as slice-level abnormality labels for the abnormalities of the fetal lungs to make the problem more tractable. We could then try to turn this into a multi-class classification problem outputting the specific type of abnormality. This will require more data and more specific labels as well.

For future direction we can view the 3D data as a sequence of 2D images and employ a sequential model to the 2D input slices using recurrent neural networks. A recent work from Monika Grewal et al. from Parallel Dots [5], 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.

Lastly, architectures designed for 3D point data have potential to efficiently process MRI data in the future. VoxNet [16] 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 [21] 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. However, this is an open problem of converting MRI scans to sparse representations.

7. Contributions

Dominic: **Preprocessing:** Made script to decompress all DICOM files and organize them in one place to be fed into model. **Dataset Statistics:** Implemented data augmentation on 2d and 3d data. Wrote metric evaluation and graphing code. Wrote scripts to assist in error analysis. **Baseline** Modified the baseline PyTorch model to handle 2D grayscaled files and binary classification. **Writeup:** Expanded dataset, augmentation, normalization, discussion and future work sections.

Mazin: Processed the compressed DICOM files to form the decompressed data set. Helped with normalization of the data set. Tried to validate and correct the orientations of each DICOM series. Tried to run open source implementions of Voxnet and Octnet models.

Shaimaa: **Data Collection:** (1) Coordinated curation of data set from Stanford Hospital Clinical Database, (2) Performed Quality control matching data labels to radiology report impression and summary findings **2D-CNN:** (1) Modified net class to use DenseNet-121 on 2D input binary classification (2) Modified data loader class to handle 3D numpy matrix files (3) Modified net class to handle 3D input for multi-slice 2D DenseNet-121 and 2D ReseNet-18 (4) Implemented F1, Precision, Recall, Accuracy metrics (5) Ran full experiment pipeline on DenseNet-121, ResNet-18 models training from scratch as well as the pre-trained models. and performed hyperparameter search. Integrated metrics and graphing into the 2D-CNN training **Dataset Statistics:** Computed histogram of gestational age. **Literature and Open Source Review:** Did the technical literature review of 2D models ResNet and DenseNet and 3D models: ResNet, segmentation models: V-Net, point cloud architectures (PointNet) to search for architectures that were feasible and promising. Performed medical literature review to define the problem and understand relevant radiologic features that characterize congenital lung abnormality and common scanning protocols used to produce the images. **Writeup:** Expanded the proposal to describe data and methods: current progress, results, future work.

David: **Preprocessing:** Wrote scripts to (1) process all decompressed DICOM files, aggregated them by their series UUID into 3D numpy matrices, and save to disk, (2) Normalized 2D normalized numpy slices, (3) Read the 2D normalized slices into the PIL data types, (4) Handle padding and truncating of scan volumes to fixed volumes. **Data Split:** Split 3D numpy matrix files into training, dev, and validation sets for both 2D volumes and 3D volumes. **Dataset Statistics:** Wrote Jupyter Notebooks to get statistics on number of slices per scan, dimensionality of each slice, and examine individual slices to validate transformations. **Baseline:** Modified baseline PyTorch DataSet class to handle 2d grayscaled fetal dataset and binary classification. Ran the full baseline pipeline with hyperparameter tuning. **3D-CNNs:** Rewrote implementation of 3D-ResNet-18 from Torch to PyTorch to process our fetal dataset. Added a DataSet class to handle 3D numpy arrays. Worked on U-Net, 3D-U-Net, and V-Net PyTorch model implementations (only to later realize that our dataset is not suited for these segmentation architectures). Ran full pipeline for 3D-CNNs and performed hyperparameter search. Integrated metrics and graphing by Shaimaa and Dominic into the 3D-CNN training script. **Literature and Open Source Review:** Took lead of technical review of literature and technical implementations for 2D and 3D segmentation models (U-Net, V-Net, 3D-UNet), point cloud architectures (PointNet, OctNet, VoxNet), and 3D-ResNet to search for architectures that were feasible and promising. Initial idea was to use VoxNet or PointNet or a segmentation architecture but differences in datasets were too large. **Writeup:** Expanded the proposal to describe data and methods: current progress, results, architecture table and details, future work.

Shaimaa, David, and Dominic: **Poster:** Contributed equally to creating and presenting the poster. David printed it out from Walgreens.

7.1. Code Base

Our public Github repository can be found here.

References

[1] 3D-ResNet Implementation in Torch, `https://github.com/kenshohara/3d-resnets/blob/master/models/resnet.lua`.

[2] C. Balassy, G. Kasprian, P. C. Brugger, M. Weber, B. Csapo, C. Mittermayer, M. Hörmann, and D. Prayer. Mri investigation of normal fetal lung maturation using signal intensities on different imaging sequences. *European radiology*, 17(3):835–842, 2007.

[3] S. Deshmukh, E. Rubesova, and R. Barth. Mr assessment of normal fetal lung volumes: a literature review. *American Journal of Roentgenology*, 194(2):W212–W217, 2010.

[4] 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.

[5] 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.

[6] 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.

[7] K. Hara, H. Kataoka, and Y. Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? *arXiv preprint arXiv:1711.09577*, 2017.

[8] 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.

[9] 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.

[10] 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.

[11] G. Kasprian, C. Balassy, P. C. Brugger, and D. Prayer. Mri of normal and pathological fetal lung development. *European Journal of Radiology*, 57(2):261–270, 2006.

[12] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] F. Milletari, N. Navab, and S.-A. Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *3D Vision (3DV), 2016 Fourth International Conference on*, pages 565–571. IEEE, 2016.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[23] E. Rubesova. Why do we need more data on mr volumetric measurements of the fetal lung? *Pediatric radiology*, 46(2):167–171, 2016.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

A.4. 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. 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.

A.4.1 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.

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.

A.4.2 Vascular abnormalities

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

A.4.3 Combined lung and vascular anomalies

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

A.5. 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.6. Bias field correction

MRI images are affected by bias field distortion, which causes the intensity to vary even across the same tissue [19]. The N4ITK method [25] is the most common method for correcting this.